

Université Paris IV – Sorbonne **École doctorale Concepts et Langages**

THÈSE

pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE PARIS IV – SORBONNE

Discipline : Informatique

présentée et soutenue publiquement le 8 décembre 2005

par

Antoine Isaac

Conception et utilisation d'ontologies pour l'indexation de documents audiovisuels

Devant le jury composé de :

Rapporteurs :	Mme Sylvie Lainé-Cruzel Mr Gilles Kassel
Examinateur :	Mme Arlette Boulogne
Encadrant INA :	Mr Bruno Bachimont
Co-directeur de thèse LaLICC :	Mr Philippe Laublet
Directeur de thèse :	Mr Jean-Pierre Desclés

Laboratoire Langages, Logiques, Informatique, Cognition, Communication, UMR CNRS 8139
Institut National de l'Audiovisuel







Mis en page avec la classe thloria.

Remerciements

Mes remerciements vont tout d'abord aux deux personnes qui ont assuré l'encadrement « quotidien » de cette thèse : Bruno Bachimont et Philippe Laublet. En plus de remplir brillamment leur fonction de pourvoyeurs d'idées, de relecteurs rigoureux, ils ont su encourager un doctorant – parfois sceptique – à aller jusqu'au bout de son effort de recherche, ce qui n'est pas rien.

Je remercie également mon directeur de thèse, Jean-Pierre Desclés, pour son ouverture d'esprit. Il m'a en effet permis de suivre des pistes qui s'écartaient des théories qu'il s'attache à défendre habituellement.

Merci ensuite à Gilles Kassel, qui a accepté de présider le jury de cette thèse. Je tiens à signaler que l'intérêt qu'il a bien voulu accorder à mes travaux, avant même l'étape du rapport, a été pour moi une grande source de motivation. Je remercie aussi Sylvie Lainé-Cruzel et Arlette Boulogne. Leur participation à mon jury a été l'occasion pour moi de confronter mes positions à celles d'expertes avisées en sciences de l'information et de la documentation, confrontation que j'estime tout à fait enrichissante.

Mes pensées vont également à l'ensemble des participants du projet Opales, avec qui j'ai découvert ce que faire de la recherche appliquée voulait dire. Avec une mention toute particulière pour l'équipe « locale » de l'INA : Véronique, Fabrice et Patrick.

Des remerciements honnêtes ne peuvent ignorer le rôle fondamental joué dans le déroulement de cette thèse par la Direction de la Recherche de l'INA, dont je tiens à remercier l'ensemble des membres, et en particulier ceux des équipes *Métadonnées* et dca dont l'activité a rythmé mon passage à l'Institut.

Pour les amateurs, voici venu le temps de remerciements moins institutionnels, avec leurs brochettes de prénoms.

Au Lalicc, je pense tout particulièrement à l'équipe de la Cuisine, et en particulier à Marie, Aude, Carine, Motasem et Jorge. Les thés et petits gâteaux ont contribué à rendre mes visites là-bas extrêmement plaisantes.

A l'ina, mes pensées vont vers ceux qui ont partagé avec moi pauses café et repas, sur aire d'autoroute ou en tente, et aussi bien avant, pendant ou après ce qu'il sera convenu d'appeler l'Age d'Or, dont les anciens se souviennent les larmes aux yeux : Jérôme, Estelle, Younès, Yohann Le Mortier, Thomas, Rémi, Jean-Calude, Matthieu, Fabienne, JPP, Benedetta, Steffen, Claude, Yann. Avec une attention spéciale pour Véronique et Raphaël, pour tout ce qui a été partagé dans le bureau des ontologues, mais aussi au dehors.

Au-delà, nombreux sont évidemment ceux que je pourrais saluer, qu'ils soient de Paris, Rouen ou Dieppe – et oui. Amis, famille, ceux d'entre eux qui auraient ce manuscrit dans les mains pourront se reconnaître.

Le lecteur extérieur ne trouvera pas ici de témoignage de gratitude envers un quelconque dieu, sauf peut-être une allusion bien mystérieuse à l'esprit du Lama Dominant à qui j'espère rester fidèle.

Il ne trouvera pas non plus d'autres prénoms, excepté évidemment celui de Cécile, à qui je dois désormais tant.

Table des matières

Table des figures

Liste des tableaux

Liste des codes sources

Introduction

1 L'INA, endroit rêvé pour effectuer une thèse CIFRE ?

L'Institut National de l'Audiovisuel – INA, ce sont d'abord, pour le grand public, des images au charme désuet[1] évoquant les temps glorieux où la radio-diffusion nationale était une. Des images que l'on voit ressurgir avec bienveillance, et même intérêt, comme de vieilles photos de famille que l'on pensait avoir définitivement perdues. De fait, et comme l'atteste l'intérêt de nombreux chercheurs pour les fonds de l'Institut, ces documents constituent bien une partie de notre histoire. Ils fournissent des témoignages précieux sur les événements, les personnages, les préoccupations qui ont été la raison de leur création et de leur diffusion. Ainsi, l'INA s'est vu reconnaître tout à fait officiellement une *mission* importante : la conservation des archives audiovisuelles produites et diffusées dans le cadre national. Il exerce également une seconde activité, plus commerciale : la revente d'extraits de documents audiovisuels aux professionnels de la production. Ces deux occupations ne ciblent pas les mêmes fonds documentaires, et ont impliqué la création de deux directions opérationnelles distinctes :

- le plus ancien est le *Département des Archives* (DDA), créé à partir du fonds d'archives de l'Office de Radiodiffusion et Télévision Française (ORTF), dont la disparition en 1974 constitue également l'acte fondateur de l'INA. Ce département a pour mission la conservation et la valorisation commerciale d'un ensemble de documents qu'il met à la disposition des journalistes, réalisateurs, documentaristes, qui peuvent ainsi se servir d'extraits choisis pour illustrer les documents qu'ils produisent.
- l'*Inathèque de France* a été quant à elle créée en 1995. Elle résulte de la volonté des pouvoirs publics d'organiser un Dépôt Légal de la radio et de la télévision indépendamment de toute préoccupation commerciale. La mission de l'Inathèque est de capter l'ensemble de la production audiovisuelle française diffusée sur les chaînes de radio et de télévision – réseaux hertziens, câble et satellite – et d'organiser la conservation de ce patrimoine et sa mise à disposition en direction d'un public sélectionné.

Cette double mission impose à l'Institut la résolution de nombreux problèmes. En effet, les fonds dont il a la garde sont gigantesques : si l'on considère l'ensemble des documents des deux départements mentionnés, on se retrouve face à deux millions d'heures de programmes, un nombre qui augmente évidemment chaque année du fait de la création et de la diffusion de nouvelles émissions. L'importance de ce volume pose bien sûr de gros problèmes de collecte, de stockage et de conservation, beaucoup d'archives étant encore sur support analogique, unique et périssable. Mais gérer l'accès à une telle quantité de documents ne va pas non plus sans difficulté : sans possibilité de retrouver efficacement ce qui est archivé, l'archivage perd de sa pertinence.

[1]Lorsqu'on entre à l'Institut, l'une des premières choses que l'on apprend est que nombre d'entre elles sont en fait menacées par les atteintes du temps et ses outrages physiques, quand elles ne se sont pas envolées en fumée avant que la main bienveillante d'un agent du Plan de Sauvegarde et de Numérisation n'ait pu les sauver d'une mort annoncée.

Comment livrer à celui qui effectue une recherche dans un fonds un document répondant à ses besoins ? Quel *traitement documentaire* effectuer ?

Pour trouver des réponses à ces questions, l'INA emploie des chercheurs[2] dont on espère qu'au contact des professionnels du monde documentaire ils sauront éviter une formalisation purement théorique des problèmes que l'on rencontre de ce monde, et des solutions qu'on peut leur apporter.

Dès le début[3], notre recherche a donc été marquée par un double ancrage : universitaire, puisque nous étions rattaché au laboratoire LaLICC[4] de l'Université Paris IV – Sorbonne, mais aussi industriel, puisqu'en tant que doctorant CIFRE (Convention Industrielle pour la Formation et la Recherche en Entreprise) nous étions considéré comme employé de plein droit de l'INA.

2 Problématique

L'objet de notre recherche est l'*indexation*. Nous verrons dans le premier chapitre de ce manuscrit que cette opération consiste à interpréter et reformuler un contenu documentaire pour le rendre accessible et exploitable par un système d'information. En fait, l'indexation des documents audiovisuels (AV) est une activité ancrée dans le quotidien de l'INA. Depuis la création de l'Institut, ses documentalistes créent des représentations qui forment la base des données par l'intermédiaire de laquelle on peut accéder au contenu des collections documentaires.

Le contexte de cette pratique est cependant en train d'évoluer. Tout d'abord, on peut observer de façon évidente la mutation de son environnement technique : l'accès informatisé aux contenus documentaires change la manière dont on peut exploiter ceux-ci, et crée des besoins nouveaux en termes d'outils et de méthodes [Bac98]. Cela induit de manière plus générale des changements du contexte organisationnel dans lequel s'effectue l'indexation : les grandes institutions doivent se poser la question de modalités d'accès différentes à leurs fonds, et on peut observer l'émergence de petites communautés travaillant de façon précise, exigeante, à partir des documents dont elles peuvent assurer elles-mêmes la production, l'indexation et l'accès. De fait, les professionnels reconnaissent que ce en quoi doit constituer l'indexation change, glissant progressivement de la synthèse – le « résumé » d'un contenu – à l'analyse – la mise en valeur des éléments du contenu qui sont pertinents pour un besoin donné – [Wal99].

Comme on le verra tout au long de ce manuscrit, notre recherche a tenté de concilier deux mouvements d'essences différentes.

Tout d'abord, deux *problèmes* fondamentaux peuvent être dégagés de l'analyse des traitements documentaires : la création des index, puis leur recherche. Accéder au contenu des documents audiovisuels nécessite en effet une indexation. Or ce processus est délicat : plutôt fastidieux à accomplir, il demande d'anticiper sur les besoins d'accès aux index construits, ce qui exige que ceux-ci aient un aspect consensuel. A l'INA, les documentalistes font appel à une solution *thésaurale*, et à leur expérience – connaissance du fonds et stratégie de recherche adaptées. Cependant, ces solutions sont, comme on le verra, encore approximatives, et des observations montrent que des raisonnements mobilisant des connaissances implicites des domaines dont re-

[2]Le Département Recherche et Expérimentation de l'Institut comporte plusieurs entités que nous ne présenterons pas ici. Celle dans laquelle nous avons effectué cette thèse est le *Groupe de Recherche Audiovisuelles et Multi-Médias* dont les travaux sont en particuliers orientés vers l'impact des techniques numérique sur le traitement des contenus audiovisuels – en ce qui concerne la restauration, par exemple – et leur description – indexation, gestion des méta-données. . .

[3]Ce manuscrit de thèse est le fruit d'un travail de recherche qui a démarré il y a de cela quatre ans, avec la réalisation d'un stage de recherche de DEA.

[4]Langages, Logiques, Informatique, Cognition, Communication, `http://www.lalic.paris4.sorbonne.fr`.

lèvent les documents seraient utiles au processus de recherche. Il serait donc bon de fournir une aide à la création et à l'accès à des index qui tiennent compte de la nécessité de précision et de raisonnement, mieux que ne le permettent les systèmes de recherche par mot-clés ou thésaurus actuels.

D'un autre côté, les *disciplines* de l'informatique, et plus particulièrement de l'intelligence artificielle, apportent des réponses à des problèmes génériques qui pourraient bien être liés aux problèmes documentaires concrets de l'Institut. Des techniques d'*ingénierie des connaissances* permettent en effet d'encoder les significations de ressources conceptuelles de manière à les rendre manipulables par des systèmes informatiques. Ces techniques ont parmi leurs objectifs d'améliorer l'accès, le partage et la réutilisation de ces connaissances entre systèmes, mais aussi entre utilisateurs humains. Leurs principales caractéristiques sont

- la formalisation des connaissances : ces dernières deviennent utilisables par l'ordinateur qui mène des raisonnements ;
- la relative richesse expressive dont on dispose dans ce cadre formel : les langages de représentation comprennent toutes sortes de concepts, relations, règles de définition et de raisonnement qui permettent d'organiser les connaissances exploitées par le système.
- la généricité des architectures envisagées : celles-ci utilisent des langages de représentation de connaissances qui sont adaptables à des domaines et des applications variables – des *ontologies* spécifient la signature du langage construit, ainsi que les règles de raisonnement spécifiques qui s'appliquent à ses éléments ;

Ces techniques, qui résultent d'efforts de recherche déjà anciens, sont actuellement popularisées par le *web sémantique*, qui essaie de les repenser pour qu'elles puissent s'adapter à des besoins réels et au passage à l'échelle induit par la dimension du web.

Notre thèse a donc dû aborder le problème de l'indexation et de son usage dans les systèmes d'un point de vue dual, se rapportant à la fois aux pratiques existantes et à des thèmes issus des disciplines scientifiques et techniques. Son enjeu est une réflexion méthodologique sur l'articulation entre ces usages et des techniques plus ou moins abstraites. Nous ne devions pas perdre de vue en effet qu'expérimenter de nouvelles pistes en matière de système d'information dans un Institut comme le nôtre, c'est proposer des instruments pour une pratique existante, mais aussi voir comment ces systèmes techniques sont susceptibles de modifier cette pratique. En l'occurrence, comme on le verra plus tard dans ce manuscrit, le cadre d'indexation que nous proposons a des répercussions sur la chaîne de description et de recherche documentaires, puisque l'on assistera désormais une partie des tâches de recherche qui se faisaient auparavant manuellement.

On peut donc synthétiser notre problème comme étant celui de l'adaptation des technologies venues de l'intelligence artificelle et de la représentation des connaissances à :

- un contexte : le document AV, que sa digitalisation met à disposition de toujours plus de personnes, avec des besoins précis et variés ;
- une pratique : l'indexation. nous mentionnons que nous nous intéressons à la fois à la création et l'exploitation des index, toutes les deux nécessaires à la constitution de bases documentaires audiovisuelles exploitables.

Dans un premier temps, ces technologies impliquent en effet une complexification des données et des traitements en eux-mêmes : les index sont structurés, et leur manipulation fait appel à des connaissances de raisonnement que l'on souhaite les plus précises possibles. Dans un deuxième temps, l'adoption de ces technologies amène une complexification de la pratique d'indexation elle-même. Nous visons en effet une indexation riche, manuelle, qui dans l'état actuel des connaissances scientifiques nécessite toujours l'expertise d'humains qui connaissent les domaines

d'application des systèmes documentaires, mais ne maîtrisent pas forcément les techniques de représentation que ceux-ci sont censés utiliser.

Il faut par conséquent s'efforcer de réduire cette complexité. Tout d'abord, on peut faire remarquer que celle-ci est liée au fait essentiel que les techniques et langages proposés dans les disciplines scientifiques que nous avons évoquées sont *formalisés.* On donne aux ressources conceptuelles qu'utilisera le système – les index aussi bien que le vocabulaire qui permet de les exprimer – une signification formelle interprétable par des systèmes d'inférences. Les sémantiques formalisées employées pour cela – comme les sémantiques ensemblistes – sont évidemment différentes de celles, non formalisées, qui donnent aux ressources traditionnelles d'une application une signification claire pour les humains qui les manipulent. Quelles que soient les théories interprétatives que l'on retient dans chacun de ces deux mondes, on est sûr de se retrouver face à une forme d'incompatibilité, ces théories ayant pour but d'exprimer des éléments de signification irréductibles les uns aux autres.

Et pourtant, nombreux sont les systèmes informatiques de gestion et de recherche d'information qui fonctionnent plutôt correctement au regard de ce que l'on attend d'eux. Lorsqu'on prend en compte les besoins d'une application de manière convenable, le recours à des *calculs* visant à assister celle-ci n'est pas nécessairement contre-productif, loin s'en faut. Ce qu'il faut, c'est concevoir les ressources du système et la manière dont celui-ci les exploite de sorte à ce qu'elles soient en accord avec les usages et compréhensions du domaine – en d'autres termes, les *contraintes sémantiques* que l'on associe aux entités que l'on y rencontre.

Dans notre cas, il faudra donc proposer une *méthodologie d'indexation à base de connaissances formelles*, dont les réflexions concerneront tout à la fois :
- la conception de ressources conceptuelles, d'*ontologies*[5] apportant le vocabulaire d'un langage d'indexation formalisé ;
- l'utilisation de ces ressources pour la création d'index formant une base de connaissances ;
- l'exploitation des index pour les recherches : les éléments du vocabulaire doivent être associés à des connaissances de raisonnement utilisables par des outils d'inférences pour améliorer les performances du système.

3 Déroulement de la thèse

3.1 Cheminement de notre recherche

Le projet Opales

Pour parvenir à accorder les nouvelles techniques avec les usages existants ou pressentis, l'INA mène une réflexion s'appuyant à la fois sur sa connaissance des pratiques documentaires, sur des travaux fondamentaux, ainsi que sur des projets expérimentaux. L'un de ceux-ci, Opales[6], a eu pour objectif la construction d'outils informatiques pour l'indexation et la recherche documentaires de contenus audiovisuels dans le cadre de communautés de pratiques extrêmement précises. En l'occurrence, pour son évaluation, il a été retenu deux *points de vue* :
- l'exploitation des documents scientifiques conçus dans le contexte de l'ethnologie, en particulier ceux qui concernent la *petite enfance* ;

[5] On verra dans le chapitre 2 une définition précise de ces objets.

[6] Opales (Outils pour des Portails AudiovisueLs Educatifs et Scientifiques) est un projet priamm. Autour de l'INA qui en a assuré la coordination, le projet regroupait le lirmm de l'Université Montpellier – II (équipes ihm et gc) et l'entreprise CS-Systèmes d'Information pour la plate-forme technique, la Maison des Sciences de l'Homme, le cnrs, le Centre National de la Documentation Pédagogique (cndp) et La Cinquième-bps pour la sélection et l'indexation de contenus vidéos. `http://opales.ina.fr`.

– l'analyse de documents pédagogiques à dominante géographique, abordant spécifiquement le thème de l'*eau*.

Pour chacun de ces points de vue, il a fallu créer un langage d'indexation, dont le vocabulaire est fourni par une ontologie. Les index sont ensuite exploités, de façon adaptée au point de vue dont il relèvent, par un mécanisme d'inférence qui assiste la recherche d'information dans la base ainsi créée[7].

Dans ce projet, nous avons tout d'abord participé à l'élaboration de la partie technique, en créant un outil de conception d'ontologies (**DOE**) intégré au système général. Ceci nous a permis de voir quels peuvent être les enjeux d'une plate-forme informatique d'indexation formelle. Nous avons ensuite collaboré très activement à la conception des ontologies liées aux deux thématiques retenues, en accord avec les besoins des experts. Finalement, nous avons été impliqué dans les séances d'évaluation du projet (création d'index et tests qualificatifs) en compagnie des dits experts. Cela nous a permis d'avoir des avis de première fraîcheur sur la pertinence des solutions techniques, mais aussi sur celle des ressources ontologiques que nous avions conçues. Et, bien évidemment, cela a contribué à nourrir les réflexions méthodologiques qui avaient présidé à la conception de ces outils et ressources.

Liens avec la recherche sur les ontologies à l'INA

La participation de notre Institut au projet Opales a de fait été l'occasion de prolonger un effort déjà ancien portant sur la conception et l'utilisation d'ontologies pour la description des contenus audiovisuels, dans la lignée des travaux de Bruno Bachimont [Bac00]. De fait, celui-ci a réuni une équipe de jeunes chercheurs dont les travaux de thèse sont complémentaires. Tout d'abord, Karine Lespinasse a abordé le thème de l'enrichissement de thesaurus par acquisition terminologique générique [Les02]. Les travaux de Véronique Malaisé utilisent des moyens similaires, mais se concentrent d'avantage sur leur application à des corpus spécialisés, en vue d'obtenir des définitions plus précises, mieux structurées, à même d'initier convenablement le processus de conception ontologique [Mal05]. Estelle Le Roux a également utilisé des outils d'analyse linguistique, mais cette fois-ci en vue de la création d'index structurés à base de connaissances, et non de celle du langage conceptuel utilisé pour représenter ces index [Le 03]. Raphaël Troncy est finalement celui dont les centres d'intérêts ont été les plus proches des nôtres : sa thèse aborde en effet le thème des bénéfices de l'emploi des ontologies pour la description documentaire. Mais elle le fait d'une manière plus concentrée sur la création de ressources spécifiquement documentaires, et son orientation méthodologique est moins globale que la nôtre en ce qui concerne la conception des ressources ontologiques elle-mêmes [Tro04], surtout si l'on considère l'intérêt que nous portons aux descriptions et connaissances de raisonnement de nature *relationnelle.*

Il reste que beaucoup de nos travaux – et des résultats de ceux-ci – ont porté sur des problèmes partagés, et ont débouché sur la réalisation d'objets communs, comme on pourra le constater lors de la lecture de notre chapitre consacré aux expérimentations que nous avons menées lors de cette thèse. Les ontologies que nous avons réalisées, les outils que nous avons construits, les réflexions méthodologiques menées ont ainsi pu bénéficier d'efforts qui, s'ils gardaient des points de vue particuliers, partageaient un objectif de recherche commun : contribuer à la définition – et démontrer les conditions de faisabilité – d'un nouveau paradigme d'indexation et de recherche documentaire.

[7] Le lecteur de ce manuscrit verra tous ces aspects plus en détails, au fur et à mesure de leur mobilisation dans le fil de notre argumentation. Il peut aussi consulter notre article décrivant le projet dans sa globalité [ICG+04].

Liens avec la recherche à l'INA et l'Institut dans son ensemble

La réalisation d'un tel objectif, même dans le cadre plus restreint qui était le nôtre, a forcément exigé une familiarisation approfondie avec le monde documentaire – ou en tout cas du sous-ensemble du monde documentaire que représente l'INA. Les paradigmes qui le structurent, les problèmes que l'on y rencontre...

Tout au long de ces trois ans, nous avons donc suivi les efforts de recherche des équipes de l'Institut qui abordaient des domaines connexes au nôtre : segmentation automatique [PC05], détection de contenu textuel dans les images [LWVJ04], construction d'interfaces élaborées de recherche [TVV04]... Toutes ces recherches ont à voir avec le document audiovisuel, ses usages, la manière dont il faut – ou dont on peut – le décrire et le rechercher.

Plus important pour nous, nous avons pu participer à des groupes de rencontre entre chercheurs et documentalistes de notre Institut. Des experts du monde documentaire nous ont ainsi présenté les outils qu'ils utilisaient (thesaurus, interface d'indexation et de recherche, etc.), la manière dont ils les utilisaient, les problèmes qu'ils rencontraient. Nous avons eu également accès aux traces les plus objectives du métier de documentaliste : les immenses bases de descriptions documentaires des départements opérationnels de l'INA, mais aussi les compte-rendus des sessions de recherche effectuées par les documentalistes dans ces bases.

Nous avons donc pu appuyer nos réflexions sur une analyse des usages dans la documentation, ainsi que sur un aperçu global des évolutions en cours et des pistes de recherche pour le plus long terme. Cette imprégnation de notre recherche par la problématique documentaire au sens large a été extrêmement utile, dans la mesure où le résultat de notre travail a la prétention d'être utilisé dans un cadre appliqué.

3.2 Retour sur ce qui a été fait dans la thèse et plan du manuscrit

Ces différents niveaux de réflexion et d'expérimentation ont permis de consolider les propositions que contiennent ce manuscrit. Le plan de celui-ci essaie de les organiser en une progression logique, du problème général posé par les pratiques existantes aux solutions précises que nous présentons pour résoudre des points particuliers de ce problème.

Tout d'abord, notre thèse apporte une réflexion sur les usages actuels en matières d'indexation et de recherche, et met le doigt sur leurs limites les plus flagrantes vis-à-vis de la problématique que nous avons dégagée. Le chapitre 1 insiste en particulier sur la définition et les enjeux de l'indexation, et la manière dont on y répond le plus souvent dans le monde documentaire. Il abordera en particulier la notion de *langage de description contrôlé*, et montrera que les techniques de spécification de vocabulaire de description les plus avancées à l'heure actuelle, les thesauri, sont loin de répondre à tous les besoins. Adaptés à un certain style de description et d'accès à des bases documentaires, ils peuvent en effet pécher par manque de précision ou de rigueur pour certaines des applications documentaires qui émergent actuellement.

Ensuite, notre thèse propose une présentation clarifiée et cironstanciée des solutions que peut apporter la discipline de l'ingénierie des connaissances aux problèmes de l'indexation documentaire. C'est l'objectif de notre chapitre 2, qui introduit les notions de *systèmes à base de connaissances* pour créer et exploiter des index structurés, et d'*ontologies* pour définir les langages de représentation et les connaissances de raisonnement qui constitueront la partie variable des spécifications de ces systèmes, celle qui fera qu'ils seront adaptés à des applications et des domaines particuliers.

Les propositions concrètes de notre travail sont présentées dans les chapitre 3 et 4 de ce manuscrit, dans lesquels nous nous sommes en effet attaché à l'obtention d'une méthodologie

globale de conception et d'utilisation des ontologies.

Nous allons d'abord présenter comment on peut guider le travail des utilisateurs devant utiliser une ontologie donnée pour créer des index. Nous reprenons les thèses de Bruno Bachimont en ce qui concerne l'accès à la *substance* des index, à savoir les concepts et relations qu'apportent une ontologie : il est indispensable de présenter une signification langagière rationnelle de ces éléments, problème que nous pouvons résoudre en utilisant des *définitions différentielles*. Nous avançons ensuite une manière de guider le travail d'un indexeur en ce qui concerne la *forme* des index proprement dite, à savoir la manière dont le vocabulaire se manifeste en une description structurée. Des *patrons de conception*, structures relationnelles adaptables qui peuvent servir de descriptions pivots vis-à-vis des procédures de raisonnement employées lors de la recherche, apportent en la matière des recommandations appropriées.

Comme nous recherchons un cadre méthodologique global, nous nous devons de nous intéresser aussi à la conception des ressources ontologiques employées par les systèmes d'indexation à base de connaissances. Nous avons proposé lors de notre thèse, et en collaboration avec Raphaël Troncy, un éditeur d'ontologies – **DOE** – qui met en œuvre les points méthodologiques proposés par Bruno Bachimont en tenant compte de l'expérience que nous avons acquise en appliquant ces points à des ontologies d'importance significative. Cet éditeur a été créé pour être complémentaire, autant que faire se peut, avec les autres outils rencontrés dans le domaine. En matière de patrons d'indexation, nous proposons d'adapter les démarches d'ingénierie ontologique à base de patrons de conception. Ces propositions souffrant à nos yeux d'un déficit trop grand en matière de légitimité applicative, nous montrons qu'il est souhaitable – et possible – de les relier explicitement aux patrons d'indexation, ces derniers étant considérés comme les patrons d'utilisation applicatifs de l'ontologie à concevoir.

Finalement, nous reviondrons dans le chapitre 5 sur les différentes expérimentations concrètes que nous avons réalisées au cours de cette thèse. Nous allons présenter brièvement les ressources ontologiques conçues, mais aussi la manière dont elles ont été utilisées, et ce que cette expérience de conception et d'utilisation a pu apporter à notre réflexion. Ce chapitre sera notamment l'occasion d'évoquer des points qui constituent à notre avis des pistes intéressantes pour des recherches à venir, ou bien tout simplement valent la peine d'être évoquées si l'on veut rendre compte de la richesse du questionnement que les problèmes d'ingénierie des connaissances peuvent apporter.

Première partie

Introduire les ontologies dans le processus d'indexation

Chapitre 1

La description des documents audiovisuels dans les processus documentaires

1.1 Introduction

1.1.1 Le statut documentaire du document audiovisuel

Affirmer aujourd'hui que le document audiovisuel tient une place importante n'a rien d'original. Et de fait, le document audiovisuel est un objet que l'on peut considérer à présent comme en cours de banalisation. Il est en tout cas de plus en plus répandu, que ce soit dans les espaces de diffusion publics (la mutiplication des chaînes de diffusion télévisuelles en est un signe), *via* des réseaux d'espaces privés mis en relation par des logiciels d'échange de données (voir l'essor récent des réseaux d'échange *peer-to-peer*), ou au sein de communautés scientifiques qui l'utilisent comme outil de travail (les laboratoires impliqués dans le projet OPALES sont dans ce cas). Et, assez souvent, on ne le considère plus que comme un objet de consommation courante, en oubliant qu'il est en fait inséré dans un cycle de vie plutôt complexe, articulé autour des étapes-clef de production, de diffusion et d'archivage [Auf00].

En amont de sa réception, il y a la production et la diffusion, étapes au cours desquelles on conçoit et réalise un programme avant de l'insérer dans un processus de publication. Ces deux étapes définissent un contexte applicatif[1] qu'on ne peut négliger : le document audiovisuel est conçu non seulement en fonction des savoir-faire et des techniques rencontrés dans le monde de l'audiovisuel, mais aussi en fonction de l'environnement dans lequel il sera diffusé. Dans le cas de l'émission télévisuelle, par exemple, le flux de diffusion produit par une chaîne peut être vu comme un type de discours – parfois publicitaire – que le public visé doit être en mesure de comprendre, ou, au moins, d'assimiler.

Viennent ensuite l'archivage et la réutilisation : comment conserver les documents ? Comment les (ré)utiliser de manière pertinente ? L'INA, de par les missions qui lui ont été données, doit considérer ces problèmes d'un point de vue archivistique, ce qui impose une réflexion sur le sens de la conservation du document. En plus de régler la délicate question du stockage matériel, il faut en effet organiser l'accès aux documents stockés, que ce soit en considérant leur valeur intrinsèque de documents-témoins ou leur possible commercialisation en tant qu'éléments à insérer dans de

[1] Par contexte applicatif, nous désignons l'ensemble – parfois très vague – des pratiques, utilisateurs et outils que l'on rencontre autour d'une application réelle.

nouveaux programmes.

Contrairement à ce qu'on pourrait penser, cette complexité qu'on reconnaît à l'analyse du document audiovisuel n'est pas propre à la mission de l'INA, qui doit gérer de larges fonds, avec le double objectif d'une exploitation commerciale fine et de l'exécution d'une mission de service public définie par un cadre légal précis (Dépôt Légal). Des problèmes similaires se posent, et peut-être avec plus d'acuité, pour des communautés de production et d'usage plus restreintes et a priori moins contraintes. La numérisation, en facilitant la mise en réseau et l'exploitation des documents, offre en effet la possibilité de faire évoluer des pratiques documentaires qui sont parfois au cœur de leur métier. Ainsi, si les utilisateurs rencontrés dans le projet Opales venaient avec des besoins moindres par rapport à ceux de l'INA en termes quantitatifs, leurs exigences étaient beaucoup plus grandes quant à la finesse du traitement documentaire. Lorsque les documents audiovisuels sont amenés à jouer un rôle très précis dans une pratique essentielle pour une communauté donnée, celle-ci a bien évidemment intérêt à ne pas prendre à la légère le problème des traitements à effectuer autour d'eux [Sto03b].

En fait, le document audiovisuel n'échappe pas à son statut de document. Comme les autres documents, on peut le considérer comme une inscription qui est motivée par les échanges sociaux qu'elle rend possible [Péd03]. Et une analyse ainsi qu'une utilisation sérieuses ne peuvent faire totalement abstraction de son contexte de production, de publication, de réception, que sa forme soit textuelle, audiovisuelle ou multimédia. Néanmoins, nous pouvons tenter de dégager des spécificités qui s'appliquent au document audiovisuel, dans le cas de l'obtention de descriptions de ce document.

Nous nous concentrerons tout spécialement sur la création et l'utilisation de descriptions du contenu du document audiovisuel pour des systèmes de recherche documentaire ou de recherche d'information[2]. En particulier, nous ne chercherons pas à analyser finement la manière dont la numérisation bouleverse les modalités pratiques de production, de stockage et d'accès au document audiovisuel (intégration de la chaîne de production, délinéarisation de l'accès au document), même s'il est acquis que ce bouleversement est à l'origine des préoccupations qui motivent les actions de recherche de l'INA, dont cette thèse fait partie.

1.1.2 La problématique de la description du contenu

Dans un environnement comme celui de l'Institut, on se rend rapidement compte qu'il est loin d'être facile d'organiser la gestion et la mise à disposition des documents audiovisuels. Après le problème immédiat de la conservation, il faut résoudre celui d'un accès raisonné aux documents et à l'information qu'ils contiennent, prélude indispensable à une réutilisation correcte de ces éléments. Il s'agit de permettre la compréhension par les usagers ou les clients du matériel qu'on leur propose, ainsi que de souligner sa valeur pour le besoin qui a motivé leur recherche documentaire – réutilisation pour une nouvelle production ou « simple » besoin d'information.

Nous verrons dans ce chapitre que la difficulté de l'accès à l'information renfermée dans un document, bien que commune à tous les types de documents, est particulièrement importante dans le domaine de l'audiovisuel. Et que ce problème, même dans une approche numérique du traitement de l'information, impose toujours une étape de description du contenu du document : on retrouve le processus de l'*indexation*, tel qu'il apparaît dans les systèmes documentaires classiques. L'*index*, utilisant traditionnellement une forme langagière, textuelle, synthétise les ren-

[2]Comme nous verrons par la suite, nous allons nous focaliser sur le problème de la *recherche d'index*. Cela nous permet pour l'instant de faire abstraction de la finalité effective du système d'information, qu'il s'agisse de retrouver des matériaux documentaires ou bien de l'information sans chercher à accéder à un support.

seignements pertinents pour l'application envisagée, et tient lieu de substitut du document dans le processus de recherche documentaire.

Nous allons cependant observer que le recours à une description textuelle ne résout pas toutes les difficultés. De gros problèmes de variabilité demeurent, que ce soit au niveau de la production des index ou de leur compréhension, et imposent le recours à des processus de *contrôle*, surtout en ce qui concerne le langage employé pour la création des descriptions ou la formulation des requêtes. Mais il faut faire attention à ne pas trop restreindre l'expressivité autorisée pour les index produits : en particulier, un langage d'indexation doit être suffisamment riche pour rendre compte des relations entre les éléments du contenu du document audiovisuel. Et comprendre que ces relations et les *structures* qu'elles induisent doivent être prises en compte par le système documentaire de manière à la fois souple et contrôlée, afin de garantir un fonctionnement optimal pour celui qui crée les descriptions et pour celui qui effectue une recherche parmi celles-ci.

Tout au long de ces réflexions, nous nous attacherons à analyser l'indexation telle qu'elle existe actuellement à l'INA, et comment elle s'insère dans le fonctionnement de la chaîne documentaire. En nous appuyant sur des exemples provenant de contextes applicatifs existants ou expérimentaux, nous montrerons que les systèmes actuels peuvent être perfectionnés, et doivent l'être, dans la mesure où les environnements technique et applicatif de l'activité documentaire évoluent.

1.2 Le problème de l'accès à l'information pour les documents audiovisuels

1.2.1 La nécessaire interprétation des documents audiovisuels

Les missions de l'INA lui imposent d'assurer un accès au document audiovisuel qui prenne en compte les deux aspects les plus importants de sa valeur informationnelle : l'objet de son contenu, et la manière dont celui-ci est exprimé. Les besoins concernant les informations de catalogage classiques – date de production, producteur... – ont certes leur importance, mais ils seront toujours exprimés en conjonction avec une demande relative au contenu des documents recherchés. Ainsi, le Département des Archives, dans le cadre de son exploitation commerciale du fonds documentaire, doit répondre à des requêtes comme celles de la figure 1.1, qui mentionnent tout à la fois des éléments de catalogage et des éléments de contenu mêlant mise en forme audiovisuelle et contenu thématique.

« images d'actualités illustrant des tempêtes violentes, où l'on peut voir la submersion des rivages »

« plans de scènes de rue à Alger, postérieurs à 1975 »

« Malraux en support film couleur »

« extrait du match Toulon/Marseille du 15/04/89 et réaction de l'arbitre, M Wurtz »

Fig. 1.1 Exemples de requêtes adressées au Département des Archives de l'INA

De fait, la recherche documentaire dans le cas des documents textuels se fait aussi principalement selon des critères de contenu [Wal99]. Ce besoin pose un problème qui n'est pas lui aussi spécifique à l'audiovisuel : celui de l'*interprétation*. En effet, le jugement de la pertinence d'un document dépend d'une interprétation de son contenu, qui va déterminer sa signification, sa valeur informationnelle pour le système documentaire auquel il appartient. Un utilisateur doit lire un document, trouver de quoi il parle, et voir en quoi il peut être utile pour son application.

Toute la difficulté dans la situation de l'audiovisuel vient du fait que le document audiovisuel ne repose pas sur un système d'assignation arbitraire de sens, comme le texte. Les éléments signifiants d'une image ou d'une piste sonore – à l'exception des éléments présentant du texte, écrit ou oral – n'ont pas de rapport prédéterminé à leur signification. Bruno Bachimont affirme que d'un point de vue sémiotique, l'image est un signe qui montre quelque chose et non un signe qui signifie quelque chose [Bac98]. Il s'agit plus d'une relation analogique avec la réalité : on a affaire à des formes perceptives qui sont susceptibles de recevoir toutes les significations que pourra leur donner l'expérience humaine, et non à des formes symboliques qui, comme les mots, ont un ou plusieurs sens proposés et articulés dans un système, puis renégociés en contexte[3]. Par exemple, là où un lecteur verra dans la séquence illustrée en figure 1.2 une séquence longue et creuse présentant un fauteuil vide plutôt quelconque, un autre pourra lire une allusion perfide à la vacance du pouvoir. Tout dépend en fait du contexte de la lecture de l'image : sa production et sa diffusion lui donnent une première finalité, et son archivage impose une grille d'interprétation particulière en fonction de l'usage de l'archive.

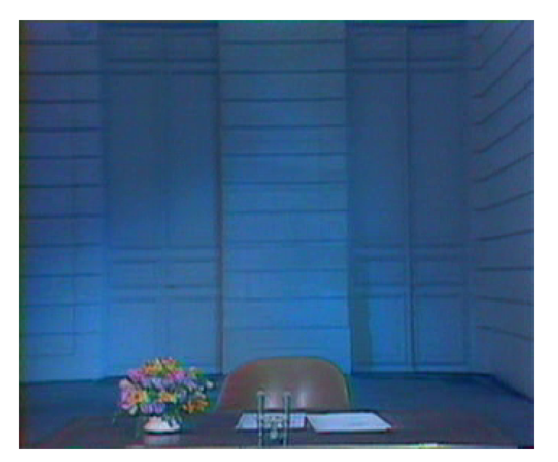



FIG. 1.2 Une séquence à la signification obscure

Un corollaire à ce problème est celui de la détermination d'une unité signifiante élémentaire : quelle est la partie de l'objet audiovisuel que l'on juge intéressante ? L'image animée propose bien des unités documentaires résultant de son découpage temporel, mais elle n'en prescrit pas une particulière comme vecteur de sens autonome. Telle analyse préférera se concentrer sur le plan, unité la plus « objective »[4] [Tur98], telle autre privilégiera la scène. Il est très délicat d'effectuer en toute généralité un traitement documentaire à un niveau de granularité plus fin que celui du document dans son ensemble. Il faut à nouveau prendre en compte le contexte de production (par exemple le genre du document) et d'utilisation (quelles sont les unités qui sont couramment recherchées et ou réutilisées par ceux qui accèdent au fonds documentaire) afin de localiser les unités dont la signification nous intéresse.

[3]Il s'agit ici d'un *contexte interprétatif* normatif, composé par le réseau des similarités et des différences de sens que présentent les unités lexicales signifiantes. C'est un tel contexte qui fait défaut aux éléments signifiants de l'audiovisuel. Néanmoins, dans le cas de l'audiovisuel comme d'autres formes documentaires, une application peut suffir à définir un environnement palliant un tel manque : le *contexte applicatif*, où ce sont les usages et pratiques mobilisées pour une application donnée qui munissent les entités manipulées d'une signification locale – et informelle.

[4]Le plan est défini comme étant le résultat d'une prise de vue continue, entre la mise en marche et l'arrêt de la caméra.

De plus, l'interprétation du document audiovisuel est rendue particulièrement délicate par le caractère intrinsèquement temporel de celui-ci. Le document audiovisuel impose en effet à son lecteur son propre rythme temporel, en dehors duquel il n'est pas complètement compréhensible [Bac98]. Pour localiser les unités significatives autant que pour déterminer leur contenu, il faut donc effectuer une lecture linéaire, complète et en temps réel, du document. La forme audiovisuelle ne présente pas de possibilité de survol, de sommaire, en bref de mode d'accès non-linéaire à ce qu'il présente, ce qui la rend encore plus difficile à traiter.

C'est pourquoi dans un système documentaire classique, pour que l'usager puisse accéder au contenu du document audiovisuel sans avoir à *lire* ledit document, un opérateur humain procède à une interprétation dont l'objectif est d'isoler les unités signifiantes du document, et de leur attribuer une signification explicite. A l'INA, il s'agit plus précisément de réduire la polysémie de ces éléments en replaçant le document dans son contexte de production, puis, selon l'optique retenue, de dégager le potentiel de réutilisation des éléments qui le composent, ou de préciser leur intérêt en tant qu'items du fonds patrimonial que constitue le Dépôt Légal de la radio et de la télévision.

On doit alors faire en sorte que cette interprétation soit correctement instrumentée. Il va s'agir d'insérer ce processus dans un cadre méthodologique et technique cohérent et efficace, assurant la pertinence du système documentaire. Le premier des problèmes est celui de l'encodage de l'interprétation : il faut en conserver une trace, pour raccourcir les recherches à venir et faciliter l'exploitation de leur résultat dans la chaîne documentaire. Nous avons observé que la nature même du document audiovisuel, et plus précisément le fait que sa signification ne rélève d'aucun système symbolique clairement défini, constitue un obstacle à son exploitation. On va voir que tout l'enjeu est d'obtenir une représentation intermédiaire satisfaisante du contenu du document audiovisuel, ce qu'on appellera un *index*. Cette représentation, qui doit être de nature symbolique afin de pallier l'indétermination sémantique du document audiovisuel et de pouvoir être manipulée correctement dans un système documentaire – informatisé ou non – servira de substitut du document dans ce système. Plutôt que de faire des recherches sur les documents, on va faire des recherches sur l'ensemble de leurs représentations.

1.2.2 La description linguistique comme substitut documentaire

Fournir et utiliser un support textuel pour la description

Dans la tradition bibliographique, qu'il s'agisse de livres ou d'images, on a recours à des *notices documentaires*, documents chargés de synthétiser les informations utiles pour la recherche documentaire : le système documentaire peut alors parcourir l'ensemble des notices afin de retrouver les documents demandés. Le tableau 1.1 donne un exemple de notice produite à l'INA. On y voit des informations de catalogage, comme le titre de l'émission et de la collection à laquelle elle appartient, le moment de sa diffusion, celui de sa création, la société qui l'a produite, les personnes qui ont participé à sa production[5]... On peut également y observer, dans les rubriques *descripteurs* et *résumé*, des informations relatives au contenu du document lui-même.

Ces notices sont de nature textuelle : elles utilisent la langue « naturelle » pour décrire le document. Et même si le « texte » obtenu n'est pas obligatoirement conforme aux règles usuelles (la notice est découpée en *champs*, qui ne constituent pas forcément un ensemble de phrases correct) il reste lisible et interprétable pour celui qui y accède, contrairement au document

[5]Les codes `REA`, `PRO`, `PRE` et `PAR` désignent respectivement le réalisateur, le producteur, les présentateurs et les participants.

Titre propre	La mort de l'infarctus
Titre collection	Comment ça va
Chaîne de diffusion	France 3
Date de diffusion	26.01.1996
Statut de diffusion	Première diffusion
Heure de diffusion	23 :18 :58
Heure de fin de diffusion	24 :10 :27
Thématique	Médecine santé
Genre	Magazine ; Reportage
Type de description	Emission composite
Médiamétrie	Culture connaissance, magazine, médecine
Générique	REA,Gauthier Bertrand ; PRO,Lanzi Jean ; PRE,Lanzi Jean ; PRE,Olivry Etienne ; PAR,Fruchart Jean-Charles ; PAR,Jordanne Jean ; PAR,Marco Jean ; PAR,Nataf Patrick ; PAR,Lemonier Sophie
Descripteurs	maladie cardio vasculaire (athérosclérose) ; infarctus ; cholestérol ; prévention ; alimentation ; régime alimentaire ; opération chirurgicale (pontage coronarien) ; artère (dilatation) ; dépistage ; recherche médicale (médecine prédictive)
Descripteurs secondaires	Toulouse ; Chine ; Lille ; Strasbourg ; Etats Unis ; ville (framingham) ; image de synthèse ; endoscopie
Résumé	Un dossier principal présente un tour du monde des pratiques alimentaires et des politiques de dépistage, avec pour guide le professeur Jean Charles FRUCHART (Institut Pasteur de Lille). Dans une deuxième partie, des reportages et des démonstrations assurées par des médecins journalistes dressent l'état des connaissances concernant la thérapie génique, les techniques de débouchage d'artères, le pontage coronarien et la prévention par une alimentation contrôlée. DOSSIER : LA MORT DE L'INFARCTUS - Après une reconstitution d'un homme faisant un infarctus, Jean Charles FRUCHART explique à l'aide d'image endoscopique comment l'athérosclérose, maladie cardio-vasculaire par excès de cholestérol, est la première cause de mortalité en France . . . DEUXIEME PARTIE - Etienne OLIVRY (médecin journaliste MVS) répond à Jean LANZI sur le rôle de la thérapie génique et de la médecine prédictive. - En duplex, le professeur Jean MARCO (clinique Pasteur, Toulouse) explique à l'aide d'animations, en quoi consiste la dilatation des artères du coeur et quelles sont les autres techniques de débouchage. - Jean LANZI et Etienne OLIVRY rappellent le principe du pontage coronarien traditionnel. - Le docteur Patrick NATAF (hôpital Pitié Salpêtrière) expose le nouveau procédé du pontage sous vidéoscopie et sur coeur battant . . .
Société de programmes	France 3
Nature de production	Coproduction
Producteurs	Producteur, Toulouse : France 3 Toulouse, 1996 ; Producteur, Levallois-Perret : Média Vidéo Son, 1996
Doc. d'accompagnement	Dossier de presse
Type de traitement	Catalogage analytique
Numéro DL	DL T 19960126 FR3 001.016
Matériel	BETA SP : 17 éléments, Parallèle antenne, Couleur, MONO, Définition : 625 lignes, Format : 1/2 pouce, Procédé : Béta, Signal : Analogique, Standard couleur : SECAM

Tab. 1.1 Extrait d'une notice documentaire

audiovisuel qu'il décrit. Le texte, bien qu'il reste soumis à des interprétations variables, a en effet recours à un système fonctionnel de signes – la langue – dont le rapport arbitraire et consensuel à un signifié permet de prescrire un sens aux unités qui le composent [RCA94]. La grande familiarité des utilisateurs avec le code linguistique achève de garantir une réinterprétation à peu près conforme à ce que le créateur de la description a voulu exprimer.

De plus, le recours à des descriptions textuelles permet d'introduire dans le système de recherche la possibilité d'une interrogation utilisant elle aussi le langage naturel. Requêtes et descriptions, se présentant sous une même forme, peuvent être plus aisément comparées en vue d'établir la pertinence des documents. Ceci est encore plus vrai dans un contexte numérique, qui permet de bénéficier des possibilités offertes par les systèmes informatiques de traitement textuel, que ce soit par le biais de la classique recherche par occurrences de mot-clefs dans le texte des notices, ou *via* des fonctionnalités plus complexes : co-occurrence, poids statistique des mots recherchés dans la notice, etc.

La forme textuelle, associée à de tels outils, permet finalement d'envisager des modes d'accès non linéaires à la description du contenu du document audiovisuel. La notice n'est pas un objet temporel ou iconique, on peut la structurer en un certain nombre de champs (voir tableau 1.1) dont les valeurs seront accessibles de manière indépendantes. On peut ainsi concentrer la recherche sur des aspects donnés en étant sûr d'accéder directement au résultat, sans souffrir des contraintes de lecture imposées par le support audiovisuel.

Notice, analyse documentaire et indexation

La notice est ce à quoi accède en premier lieu l'utilisateur lors de la recherche documentaire. A ce titre, elle doit pouvoir fournir le maximum d'informations qui puissent aider son lecteur à juger de la pertinence du document représenté.

A l'INA, on retrouve plusieurs niveaux de description, suivant l'importance accordée au document dans la politique d'analyse du fonds documentaire. Pour simplifier, et se ramener à un grille générique adaptable aux pratiques couramment rencontrées dans le monde documentaire, on peut retenir trois niveaux : l'*identification*, le *catalogage* et l'*indexation* [Pic96]. L'identification concerne les informations permettant une gestion basique du document : son titre, son producteur, les données de programmation, etc. Le catalogage développe les informations factuelles de l'identification, en y ajoutant selon le niveau de détail retenu les auteurs et participants principaux du documents (les informations de générique), la thématique générale du document ou encore un résumé « objectif » de l'enchaînement des séquences qu'il comporte. Seul le troisième niveau de description fait intervenir une phase d'interprétation. L'indexation vise en effet à catégoriser le document d'après son contenu, suivant le point de vue applicatif retenu.

Catalogage et indexation sont les produits de l'*analyse documentaire*, ce que Waller dans [Wal99] décrit comme

> *l'étape de la description qui consiste à présenter sous une forme organisée, concise et précise les données [...] contenues dans un document ou un ensemble de documents.*

L'analyse documentaire distingue les informations d'identification et de catalogage de bas niveau de celles se rapportant à un examen plus approfondi du document : la description du contenu. Cette dernière s'articule le plus souvent autour de deux produits : le *résumé* et l'*indexation*. Le premier est censé donner une représentation de ce que présente le document, fidèle à l'organisation créée par son auteur – c'est en cela qu'on a pu le qualifier d'objectif dans le paragraphe précédent. La seconde, quant à elle, a aussi vocation à représenter le contenu du document, mais elle le fait

d'un point de vue particulier, celui de l'application envisagée pour le système documentaire. C'est donc celle-ci qui va nous intéresser en premier lieu. Selon l'AFNOR [AFN93],

> *l'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse. [...] La finalité de l'indexation est de permettre une recherche efficace des informations présentes dans un fonds de documents et d'indiquer, sous une forme concise, la teneur d'un document.*

Pour Bruno Bachimont [Bac98],

> *une indexation est la paraphrase d'un contenu en une forme sémiotique permettant de rendre exploitable le contenu indexé pour une pratique donnée.*

Ces deux définitions rendent bien compte de ce qu'est l'indexation :

- une *représentation* : l'index doit représenter correctement le contenu du document analysé. Il doit prendre une forme inspirée par les usages et besoins documentaires, propice à le rendre aisément interprétable par celui qui y accède ;
- une représentation *centrée sur un usage* plutôt que sur la description en elle-même : le premier critère de pertinence d'un index est qu'il doit servir à retrouver l'information utile pour l'exploitation visée. Plus qu'à un hypothétique contenu objectif, l'indexation s'intéresse à la *signification* de ce contenu dans un cadre applicatif précis.

L'indexation se décompose de manière traditionnelle en *analyse* et *reformulation*. L'indexeur doit tout d'abord isoler les éléments pertinents du contenu du document, puis condenser ces informations sous la forme d'une description adaptée au système documentaire. Dans le cas des systèmes textuels, l'usage le plus courant est le recours aux *mots-clefs* : les aspects du document reconnus comme intéressants sont alors identifiés par une liste de mots issus du langage naturel. Leur choix, toujours délicat, dépend étroitement du savoir-faire documentaire de celui qui indexe, ainsi que de sa connaissance de l'exploitation future de l'index et du document auquel il se rapporte.

Les indexations des documents audiovisuels

Dans le cas des documents audiovisuels de l'INA, la création des index obéit à des modèles de description issus de l'observation experte des besoins exprimés par les clients-utilisateurs, et de l'anticipation des requêtes auxquelles pourraient répondre les documents [Pic96]. Ces besoins dégagés, on l'a déjà évoqué, concernent tout d'abord le contenu proprement dit du document, ce dont il traite. L'*indexation de contenu* est censée répondre aux interrogations classiques structurant l'analyse documentaire [Wal99] : Qui ? Quoi ? Où ? Quand ? Pourquoi ? .

Pour un but relativement proche du nôtre, à savoir l'analyse de documents audiovisuels élaborés mélangeant interviews et extraits, [Sto03a], [Mou99] et [LA83] soulignent bien l'importance de l'analyse du contenu : personnes, thèmes et lieux sont à mentionner. Mais elles évoquent aussi un autre impératif : la forme audiovisuelle influençant la perception et donc la compréhension de ce qui est montré, il faut en tenir compte lors de l'indexation. De fait, de nombreuses recherches documentaires ciblent plus l'objet audiovisuel en tant que support d'une information que cette information elle-même, que l'on peut d'ailleurs souvent atteindre dans des documents plus faciles à manipuler que les documents audiovisuels... On se retrouve dans la situation d'un usage analogue à celle de la citation dans le texte : il s'agit tout autant de rendre compte d'une

présentation particulière, de s'appuyer sur une légitimité donnée – pour les images d'archives – que de traiter un thème donné.

Michel Dauzats, de l'INA, faisant un état des lieux des langages documentaires dédiés à l'image dans [Dau94], rappelle bien que l'un des objectifs en la matière, a fortiori pour les documents audiovisuels, est d'indexer la forme et la technique. Ceci impose de produire des descriptions relatives aux procédés audiovisuels employés, de procéder à une analyse de la bande image et de la bande son qui peut suivre une approche morphologique ou esthétique – à la manière de Zettl dans [Zet02] – et qui demande à l'indexeur une grande part d'interprétation. Et, puisque la mise en forme peut s'appuyer sur la richesse structurelle qu'autorise le montage audiovisuel, l'analyse doit aussi rendre compte de la structure du contenu documentaire, en n'omettant pas de s'appliquer également aux séquences (des extraits d'autres films, par exemple) que celui-ci contient. A l'INA, c'est l'*indexation formelle* qui a pour rôle de présenter l'organisation, la structure des documents, ainsi que les dispositifs techniques utilisés, la répartition du contenu entre les niveaux visuel et sonore, etc.

Ces informations sont capitales pour cerner l'intérêt de l'élaboration de chaque document ou segment documentaire, et donc pour mieux appréhender l'éventail de ses utilisations pertinentes. Ce sont elles qui guident la production des notices, qui se situent de fait à plusieurs niveaux documentaires : la collection de documents (« fiches collection »), le programme (notice relative à une émission), et, quand le programme est composé et présente un intérêt suffisant, les séquences ou sous-programmes qui composent ce programme. Cette situation conduit Gwendal Auffret [Auf00] à analyser l'indexation comme le résultat de trois phases : la *localisation*, qui consiste à repérer les segments documentaires jugés intéressants, la *qualification*, qui indique ce en quoi ils sont pertinents, que ce soit au niveau de la substance même du contenu ou de sa mise en forme, et la *structuration*, qui relie les éléments décrits dans une structure méta-documentaire rendant compte de la complexité structurelle du document audiovisuel.

Finalement, une notice telle que celle du tableau 1.1 va contenir des champs relatifs à la fiche signalétique du document (les données de l'identification), des champs consacrés à la description du document, d'autres consacrés à la gestion du support matériel du document, et enfin ceux qui concernent la gestion de la notice dans le système documentaire.

Les champs les plus intéressants au niveau documentaire sont les champs relevant du contenu : *résumé, séquences, descripteurs thématiques, descripteurs image*, etc. Ce sont eux qui bénéficient de l'effort de documentation le plus important, et vont en particulier rendre compte des aspects évoqués ci-dessus, ceux qui nécessitent le plus l'interprétation du document par l'indexeur.

On retrouve cette distinction entre les données interprétatives et les autres dans le projet Opales. Que ce soit pour l'application relative aux documents sur la petite enfance ou à celle concernant l'analyse des programmes documentaires sur l'eau, les informations reconnues comme les plus pertinentes et, partant, qui ont fait l'objet de la plus grande attention, sont celles qui relèvent de la description du contenu, surtout si celle-ci, se présentant alors comme une indexation, est effectuée en fonction des besoins dégagés par la communauté qui utilise les documents.

Numérisation, audiovisuel et indexation automatique

Ce survol de l'activité d'indexation des documents audiovisuels montre bien qu'il s'agit d'un processus qui requiert une interprétation fine, s'appuyant sur une lecture approfondie des dits documents. Par conséquent, les moyens mis en œuvre doivent être conséquents, tant en ce qui concerne leur qualité que leur quantité. Indexer à un niveau correct revient cher, et de nombreuses personnes se demandent s'il n'est pas possible, à l'heure où de plus en plus de tâches semblent

pouvoir être automatisées, de faire accomplir tout ou partie du processus d'indexation par des logiciels.

Dans le cadre de la convergence numérique, à savoir celui de l'insertion des moyens informatiques dans toutes les tâches de la chaîne documentaire, de la production à l'archivage, on espère lever bien des problèmes en faisant manipuler le document audiovisuel par les ordinateurs. Cela n'est cependant pas encore applicable au problème spécifique de l'indexation. En l'état actuel des connaissances et des systèmes, il est en effet extrêmement difficile de faire en sorte que les systèmes d'indexation automatique puissent produire autre chose que des descripteurs de très bas niveau du contenu des documents : couleurs, formes, mouvements, etc. Contrairement à ce qui est souvent avancé, cette *détection de contenu* renvoie plus aux caractéristiques basiques de ce qui est montré dans le document qu'à une véritable *analyse* de ce que renferme le document, ce qu'est l'indexation[6]. Il est donc improbable que de tels moyens garantissent autres chose qu'un accès limité aux documents, puisque les informations qu'ils produisent sont très éloignées du niveau auquel sont habituellement formulées les requêtes. Les spécialistes de ce domaine reconnaissent l'existence d'un *fossé sémantique* [ES03] entre les traitements que les ordinateurs peuvent produire, et ce que les utilisateurs attendent des systèmes documentaires.

Certains outils peuvent être néanmoins utiles pour assister le processus d'indexation – segmentation automatique en plans, reconnaissance du langage parlé, reconnaissance de visages et de grosseurs de plan – ou celui de recherche – recherche d'images par similarité graphique. De plus en plus nombreux sont les chercheurs qui proposent même, pour des applications précises, d'extraire des documents des caractéristiques « sémantiques » de plus haut niveau, pourvu qu'ils puissent utiliser des connaissances *a priori* [Car00, Ven02]. Mais aucun ne peut encore être en mesure de rivaliser avec la finesse de l'interprétation humaine, surtout si celle-ci est fortement influencée par une finalité applicative particulière. Dans le projet Opales, par exemple, il s'agissait de produire des index dont le niveau de détail est hors d'atteinte pour un outil automatique : comment distinguer la mère d'un enfant de toutes les femmes s'affairant autour de lui ?

D'aucuns peuvent affirmer que la présence à côté d'un document audiovisuel de documents plus facilement manipulables, comme des documents textuels – on parle de *péritextes* [LHB00] – peut remédier à ce problème et permet d'envisager l'abandon de l'interprétation humaine. [Les02, Le 03] ont montré que de telles sources, associées aux bons outils d'extraction d'informations peuvent constituer une assistance capitale à l'indexation. Mais il n'est pas certain que ces documents existent ou soient accessibles pour n'importe quelle application. De plus, le lien entre les informations extraites des textes et les éléments de l'image (séquences temporelles, zones spatiales, objets, personnages) reste délicat à établir. Finalement, la primauté du document audiovisuel demeure, et ce pour une raison dont la simplicité confine à la tautologie : s'il était équivalent à un ensemble de documents textuels qui existent indépendamment de lui, il n'y aurait pas d'intérêt à le produire. Et de fait, les péritextes, en tant que textes, doivent obligatoirement relever d'un environnement interprétatif donné, qui n'est pas forcément celui retenu pour l'exploitation des documents audiovisuels.

[6]Bruno Bachimont rappelle dans [Bac04a] que ce que la détection automatique produit, ce sont des descripteurs *physiques*, directement liés à la perception du contenu documentaire. Alors que l'indexation vise à l'obtention de descripteurs *sémantiques*, associés à une interprétation. De fait, il n'y a pas forcément de lien entre les deux types de données, puisqu'on peut modifier un aspect physique du document – changement de grosseur de plan, variation de couleur, encadrement et logos – sans changer nécessairement la signification de ce document – ce en quoi la partie modifée était intéressante.

1.2.3 Améliorer la qualité de l'indexation dans son contexte de production et d'usage

L'indexation telle que nous la concevons restitue donc l'interprétation qu'a fait un opérateur humain de ce qu'il a lu. On doit considérer – en tout cas, c'est ce qui est fait à l'INA – qu'elle est ce qui rend le mieux compte des particularités de chaque document en le plaçant dans son contexte d'utilisation. Dans un tel cadre, il est tout de même permis de rechercher des améliorations pour ce processus, même si on ne peut pas tout automatiser. Quels sont les facteurs qui permettent de garantir, voire d'améliorer, la qualité du système de recherche documentaire ?

Si on analyse la situation en termes de chaîne de traitement et de réception de l'information, on peut distinguer trois « leviers » de changement :

- la *source de l'interprétation*. On peut orienter la manière dont l'interprétation est produite en guidant ceux-là même qui la produisent. A l'INA, des documentalistes spécialisés dans l'analyse des documents audiovisuels sont formés afin de produire les descriptions les mieux adaptées aux besoins de l'institut. Leur expérience du milieu de la recherche documentaire audiovisuelle, ainsi que la formalisation au sein de guides d'un certain nombre de règles relatives aux index produits, garantit la pertinence de leur analyse. Dans des applications plus restreintes, l'expérience montre que ce sont souvent les experts de l'application qui sont mobilisés en vue de la description des documents.
- le *support de l'interprétation*, et le comportement du *système d'information* exploitant ce support. Il faut définir l'encodage de l'interprétation en fonction de l'utilisation que l'on veut faire de cette interprétation par un système de traitement de données. La notice textuelle représente une première avancée pour l'insertion d'une interprétation pertinente du document audiovisuel dans le système documentaire qui les exploite, surtout si celui-ci bénéficie de fonctionnalités de recherche textuelle.
- le *destinataire de l'interprétation*. L'utilisateur du système peut-il être assisté en fonction de ses pratiques lorsqu'il se retrouve face à ce que l'indexeur a créé ? A l'INA, il faut distinguer deux usages. D'une part, celui de l'Inathèque de France, où le public qui accède au document n'a pas a priori une connaissance approfondie de la documentation audiovisuelle. Souvent les utilisateurs ne maîtrisent pas les concepts qui leur permettraient de rechercher correctement les informations, et sont obligés de recourir à l'aide d'un documentaliste. D'autre part, celui de l'exploitation commerciale des archives, où la médiation opérée par les documentalistes entre le système et le client devient totale : le client apporte sa requête, et c'est le documentaliste qui va rechercher à sa place le document, et surmonter les difficultés d'accès qui pourraient survenir[7].

Dans le cadre de notre thèse d'informatique, et plus largement du thème de recherche *Description des Contenus Audiovisuels*, équipe dans laquelle ces travaux ont été menés, l'amélioration doit porter sur le second item de la chaîne. Comment faire en sorte de mieux capter l'interprétation pertinente du contenu du document audiovisuel ? Comment obtenir un système qui utilise efficacement ce contenu afin de faciliter les tâches de description et de recherche documentaires ? Ces questions nous amènent au problème du *contrôle* du support et des processus de l'indexation. Le fait est qu'en restreignant les possibilités de saisie des interprétations dans un système, on peut influencer l'ensemble de la chaîne : l'indexeur ne peut pas insérer une interprétation par trop incohérente ou inadaptée, et celui qui accède à la description peut trouver une aide dans la

[7]Mais cette situation est amenée à évoluer très rapidement : la Direction des Archives a par exemple créé une interface d'accès aux documents et aux notices, INAMEDIA, qui a pour vocation d'être directement employée par le public.

manière dont elle est formulée. Ce problème a déjà été abordé par les approches documentaires : nous allons donc commencer par évoquer les difficultés que l'on peut rencontrer dans les systèmes textuels existants, et observer la manière dont elles sont gérées.

1.3 Le contrôle du support de l'indexation

La matière de l'indexation est ce que l'indexeur juge pertinent de communiquer à celui qui effectue une recherche. Il est donc important de garantir une continuité d'interprétation et d'exploitation – [DB00] parle de *continuité sémantique* – entre ce que choisit d'exprimer un indexeur et ce que va comprendre celui qui va se retrouver face à la notice. La qualité de la restitution de cette interprétation conditionne la qualité de l'ensemble du système de recherche documentaire. En effet, cette continuité concerne également, au niveau du fonctionnement du système, le rapprochement entre les descriptions produites et les requêtes effectuées. Comment s'assurer que deux éléments rapprochés l'ont été pour un motif correct, c'est-à-dire qu'ils correspondent à deux interprétations qui peuvent être mises en correspondance ?

1.3.1 Index, recherche et variabilité textuelle

Dans cette optique, le recours à la solution textuelle constitue un premier progrès : d'un document audiovisuel aux significations presque illimitées, on passe à une représentation qui s'inscrit dans un système beaucoup plus normé, où une première interprétation pourra être restituée à l'utilisateur et servira de support à sa recherche documentaire. Néanmoins, l'expérience dans le monde documentaire a vite montré que l'utilisation du texte seul, sans contrainte, ne suffit pas à garantir une continuité interprétative suffisante [Wal99]. Cela est dû à la nature même du langage, qui est propice à une variabilité importante : une information peut s'exprimer de plusieurs façons différentes, et une même expression peut trouver plusieurs interprétations. La synonymie et la polysémie sont identifiées depuis longtemps comme des facteurs conduisant à une perte d'efficacité des systèmes de recherche simples à base de mots-clefs. Il se peut qu'une notice documentaire ne soit pas incluse dans la liste des réponses données à une requête, parce qu'elle ne contient pas les bons mots, alors que son contenu est pertinent. On observe alors un phénomène de *silence*[8]. Dans le cas inverse, celui du *bruit*[9], une description non pertinente sera retournée parce qu'elle contenait les mots recherchés, mais avec un sens différent.

Les variations sont encore plus grandes lorsqu'il s'agit de créer des descriptions textuelles rendant compte de contenus riches, articulés, ce qui est souvent le cas des documents audiovisuels. Comment être sûr qu'une requête exprimée sous une forme telle que « plans de scènes de rue à Alger » puisse trouver une réponse si on considère qu'il est tout à fait possible et légitime qu'un indexeur ait pu décrire de tels plans, dans un reportage traitant de la vie algérienne, avec l'expression « images d'un marché algérois » ? Les modalités d'expression ainsi que la matière du contenu de l'objet audiovisuel peuvent être à chaque fois décrites par un très grand nombre de variantes. Ces possibilités sont si nombreuses que même la cohérence de l'indexation n'est pas garantie : différents index pourront être produits pour un même document par deux indexeurs différents, ou par le même indexeur à deux moments différents. Ces phénomènes sont de nature à diminuer l'efficacité de l'utilisation des index lors de la recherche. Si l'on veut conserver un processus de description textuelle, il faut donc adresser une réponse satisfaisante à ce problème de variabilité.

[8]Dans le domaine de la recherche d'information, le silence désigne les résultats répondant à une requête, mais non renvoyés par le système de recherche.

[9]Le bruit désigne lui les résultats faux renvoyés par le système.

A l'INA comme dans d'autres organismes, la continuité est en partie assurée par l'utilisation d'un certain nombre de savoirs relatifs à la production des descriptions. Un ensemble de connaissances, explicites ou tacites, est en effet partagé par les documentalistes tout au long de la chaîne documentaire. Un thesaurus (cf. section 1.3.2) fournit un vocabulaire commun de description, des guides de référence et des règles d'indexation renseignent sur les aspects des documents à décrire, et sur la manière de les décrire. On peut aussi essayer de tenir compte des variations historiques des indexations produites. Mais très souvent on se repose encore sur les savoirs implicites qui constituent l'environnement des documentalistes, sur leur connaissance du fonds documentaire de l'Institut, sur leur savoir-faire ou leur intuition pour répondre à des questions en compensant les effets de la variabilité. Et dès que les moyens et la nature de la mission assurée par l'organisme le permettent, on préfère laisser la tâche de la recherche entre les mains de ceux qui connaissent les arcanes du système plutôt qu'au large public des usagers et des clients.

Au cours d'observations que nous avons menées sur des sessions de recherche documentaire à l'INA, par le biais de fichiers de trace, nous avons pu analyser comment les documentalistes procèdent à une série de reformulations de leurs requêtes, au fil de leur intuition et de leurs connaissances sur le sujet concerné. Par exemple, pour une requête demandant des images sur les « débuts des supermarchés Leclerc », le documentaliste pose une requête comportant les mots-clefs « supermarché » et « Leclerc », qu'il affine suivant des critères de date ou géographique. Ces raffinements peuvent à leur tour être suivis d'élargissements ou des reformulations faisant intervenir des termes liés – de « supermarché » à « magasin » puis à « hypermarché ». Il est à noter que le dernier terme n'existe pas encore à l'époque ciblée, mais une indexation postérieure, elle-même réalisée sur un document récent, peut très bien l'avoir introduit dans le parcours qui mène au bon document. Le tableau 1.2 donne les différentes requêtes rentrées dans le système par le recherchiste ainsi « pisté ».

```
FIND <IMAGO.BASINA>emission WHERE ITOUSTEXT PH WORDS 'leclerc '&'supermar*'
```

recherche de notices comportant les mots « Leclerc » et « supermarché »
(un joker est utilisé pour compenser les variations de graphie)

```
FIND <IMAGO.BASINA>emission WHERE ITOUSTEXT PH WORDS 'leclerc'&'hypermarch*'
FIND <IMAGO.BASINA>emission WHERE ITOUSTEXT PH WORDS 'lecler*'&'magasin*'
FIND <IMAGO.BASINA>emission WHERE ITOUSTEXT PH WORDS 'lecler*'&'surface*'
```

utilisation de notions conceptuellement proches de « supermarché »

```
FIND <IMAGO.BASINA>emission WHERE ITOUSTEXT PH WORDS 'landern*'
FIND <IMAGO.BASINA>emission WHERE ITOUSTEXT PH WORDS 'superm*'&'quimper'
```

introduction de critères de localisation : le premier magasin Leclerc a été ouvert à Landerneau

```
FIND <IMAGO.BASINA>emission WHERE RDATDIF = '01.01.1965' :'31.12.1965'
 AND ITOUSTEXT PH WORDS 'quimper'
```

introduction d'une restriction de date, peut-être pour avoir des images illustratives de la ville de Quimper à l'époque de l'essor des magasins Leclerc

Tab. 1.2 Une série de reformulations de requêtes documentaires

Cependant, le monde documentaire préfère employer des moyens plus formels et donc plus sûrs de superviser la production des index. Cela paraît d'autant plus souhaitable à l'heure où de nouveaux usages se profilent, plus ouverts au monde extérieur, ce que laisse envisager la mise à

disposition directe des fonds de l'Inathèque en direction des chercheurs depuis quelques années. Le premier de ces moyens est celui du contrôle du vocabulaire employé dans les descriptions.

1.3.2 Contrôler le vocabulaire des index

La définition de l'indexation, on l'a entr'aperçu, prescrit l'utilisation d'un *langage documentaire*. Ce langage peut évidemment être le langage naturel, mais dans bien des cas on préfère utiliser un langage normé, qui permettra aux index de prescrire plus précisément une interprétation du contenu du document.

Les langages documentaires

L'usage (cf. [Wal99]) distingue les langages documentaires, affectés à la recherche par sujets, des *listes d'autorités*, qui concernent surtout les champs documentaires où apparaissent des noms propres ou des fonctions que peuvent tenir ces personnes[10]. Les listes d'autorités sont de simples catalogues de mots, dont le but est souvent de fournir une référence à l'utilisateur qui cherche à vérifier la possibilité de l'emploi d'un terme, ou d'une graphie donnée pour un nom propre ou une abréviation. Les langages documentaires sont eux définis comme des

> *langages artificiels constitués de représentations de notions et de relations entre ces notions, destinés, dans un système documentaire, à formaliser les données contenues dans les documents et dans les demandes des utilisateurs.* [AFN87]

Parmi ces langages documentaires, on trouve des langages classificatoires, pré-coordonnés : les unités documentaires sont les *sujets* que peuvent traiter ces documents, sujets qui sont présentés sous forme de codes et/ou de *vedettes-matières* rassemblés dans une arborescence fixe, du plus général au plus particulier. Le problème est que l'utilisation de cette simple arborescence, si elle permet une classification des documents en ensembles liés par la relation d'inclusion, est trop rigide dès que l'on voudrait considérer des liens « transversaux » entre sujets et introduire une certaine souplesse dans l'indexation. Par exemple, si le sujet `opération cardiaque` apparaîtra naturellement sous le sujet `opération chirurgicale`, le lien avec le sujet `coeur` apparaissant sous `organe` sera plus difficile à introduire. Il faudra systématiquement rattacher les documents concernés aux deux vedettes, ce qui peut poser problème dans une stratégie d'indexation par sujet : il y a en effet une différence fondamentale entre affirmer qu'un document traite d'un sujet, même complexe, et dire qu'il aborde deux sujets. Ou alors, si l'on est prêt à faire évoluer le langage à chaque fois que l'on rencontre une relation jugée intéressante, il faudra introduire un nouveau sujet, et le lier à ceux qui en sont proches. Cette option, si elle a l'avantage d'encoder de manière relativement claire la façon dont des notions sont articulées au sein d'un même sujet, peut rapidement faire exploser la taille du vocabulaire ...

Une grande quantité de chaînes documentaires utilisent donc des versions améliorées de ces langages, dites *combinatoires*. Dans ces langages, l'unité retenue est la *notion*, elle-même exprimée à l'aide de termes issus du vocabulaire des utilisateurs. L'idée est de faciliter la *combinaison* de ces notions pour former des sujets adaptés à ce que l'on a besoin d'exprimer, que ce soit pour l'indexation ou la recherche. A la différence de l'approche précédente, le sujet n'est plus conçu de manière statique, et l'attention se concentre sur les composants thématiques – les notions – qui permettent de le construire.

[10]L'INA par exemple propose une liste des rôles que peuvent tenir les personnes impliquées dans la conception d'un document audiovisuel : réalisateur, présentateur, producteur, etc.

Le thesaurus, instrument de contrôle du vocabulaire par excellence

L'exemple le plus abouti de ces langages est le *thesaurus*. La norme ISO 5694-1 définit cet objet comme un

> *vocabulaire d'un langage d'indexation contrôlé, organisé formellement de façon à expliciter les relations a priori entre les notions (par exemple les relations générique-spécifique).*

Les relations que l'on trouve dans un thesaurus pour organiser les termes ont comme objectif de lever les ambiguïtés du langage lors de la formulation de l'index ou de la requêtes. Elles définissent en effet un *réseau sémantique* qui, outre le fait de donner une liste des termes autorisés, apporte un contexte d'interprétation qui facilite leur compréhension et leur manipulation. Ces relations sont :

- la *relation hiérarchique* qui aide à structurer les champs sémantiques des notions en précisant les rapports de subordination entre notions, de la plus générique à la plus spécifique (cf. tableau 1.3). Elle permet par exemple de retrouver un document traitant de notions spécifiques à partir d'une question portant sur une notion plus générale ;
- la relation d'association (« *voir aussi* », « *CF* » pour *confère*) qui met en rapport des notions de sens ou d'usage voisins, situés dans des champs sémantiques différents. Elle assouplit l'organisation hiérarchique en permettant d'élargir les requêtes vers des sujets différents mais rapprochés par la pratique.
- la relation d'utilisation préférentielle (« *utilisé pour* », « *UP* »), qui en plus d'indiquer les termes dont les significations sont jugées équivalentes permet de distinguer les termes préférés de ceux dont l'emploi est prohibé. Il existe une version non orientée de cette relation, l'« équivalence », qui se contente de signaler que les termes liés sont interchangeables.

Un exemple d'organisation de termes dans un thesaurus est donné par l'extrait du thesaurus de l'INA[11] présenté en tableau 1.3, où l'indentation dénote la relation de spécialisation entre notions.

Le thesaurus est donc en mesure de faciliter la compréhension des termes dans le contexte de l'application, puisqu'il utilise un vocabulaire adapté, tout en organisant ce vocabulaire en un réseau pouvant guider l'interprétation de manière convenable. Par exemple, dans l'extrait du tableau 1.3, le terme `technique audiovisuelle` est précisé par les termes qui le spécialisent : il s'agit de procédés qui sont employés lors de la production et la transmission de l'image et du son – `accéléré`, `effets spéciaux`, `mixage` y figurent – mais pas pour l'encodage physique du document sur un support audiovisuel particulier – on ne trouve pas `16mm` ou `compression MPEG`. Au cas où il resterait des ambiguïtés, un thesaurus peut associer à ses termes des notes d'application (`NA`) qui clarifient l'emploi d'un descripteur : `image sonore` est ainsi précisé par « `Son caractéristique à fort pouvoir évocateur (événement, lieu, période, etc.)` ».

Intérêts et limites des thesaurus pour les systèmes documentaires

Les relations entre termes peuvent aussi contribuer au bon fonctionnement du système : d'une certaine manière, le thesaurus permet au système de mieux interpréter lui aussi les notions, et de les manipuler d'une manière qui soit pertinente pour l'application visée. Concrètement, le système peut aider l'appariement entre requêtes et index par le biais de reformulations en utilisant les

[11] Il est à noter que le thesaurus de l'INA est un cas particulier. Alors que la plupart des thesaurus se restreignent à un domaine applicatif donné, celui de l'INA, pour ce même impératif applicatif, est condamné à aborder tous les domaines que peuvent traiter les documents audiovisuels, ce qui lui donne une dimension encyclopédique : plus de 12000 descripteurs organisés en plus d'une centaine de champs, sur 8 niveaux de profondeur.

```
audiovisuel
       radio
              matériel radiophonique
                     poste à transistors
                     CF : PHYSIQUE/TRANSISTOR
                     poste de radio
                     UP :récepteur radio
              émission radiophonique
                     image sonore
                     NA : Son caractéristique à fort pouvoir évocateur
                     (événement, lieu, période, etc.)
                     jeu radiophonique
                     journal parlé
       télévision
       ...
       technique audiovisuelle
              animation
              CF : DESSIN ANIME
              effet d'image
              UP : effet-image
              image de synthèse
              CF : IMAGE VIRTUELLE
              montage
```

TAB. 1.3 Un extrait du thesaurus de l'INA

chemins relationnels, et ainsi soulager une partie de l'effort consistant à pallier une rupture de continuité interprétative entre l'indexeur et celui qui effectue la recherche (cf. exemple du tableau 1.2). Par exemple, si un utilisateur recherche des documents utilisant une `technique audiovisuelle` particulière, les documents dont les index contiennent la notion `animation`, liée à la précédente par la relation hiérarchique, pourront être renvoyés.

De fait, la plupart des systèmes utilisant des thesaurus exploitent les informations de spécialisation ou de généralisation pour améliorer leurs performances, surtout en termes de *rappel*[12]. L'expérience montre que cela est efficace, surtout dans le sens descendant – on fait une recherche concernant des termes plus spécialisés pour répondre à une requête employant des termes plus génériques. L'INA compte parmi ceux qui utilisent une telle fonctionnalité dans leur système de recherche. Cependant, il ne faut pas perdre de vue que la possibilité de perte en *précision*[13] est importante. Le champ `audiovisuel` contient des notions telles que `publicité` qui, si les documents correspondants étaient renvoyés, pollueraient les résultats d'une requête sur une `technique audiovisuelle`.

Certains auteurs défendent aussi la pertinence de l'utilisation des relations transversales d'association [TAJ01]. Si elles jouent un rôle moins structurant pour les champs sémantiques, elles rendent souvent compte de pratiques souvent utiles en ce qui concerne la recherche d'information dans le domaine concerné. Ainsi, pour rester dans les limites de notre exemple, un chercheur consultant le thesaurus pourra décider de remplacer dans une requête le terme `animation` par `dessin animé`, ce qui semble naturel vu le lien fort, définitoire, entre ces deux notions : un dessin animé utilise forcément une technique d'animation.

Le thesaurus propose donc un moyen efficace et relativement simple – il utilise des termes connus des utilisateurs, et leur prescrit naturellement une interprétation en restituant leur contexte d'interprétation – de *normer le vocabulaire des index* de manière à assurer la continuité de leur interprétation tout au long de la chaîne documentaire. Dans les situations simples où l'index associé au document ne contient pas toutes les notions qu'on peut légitimement associer à ses éléments, le système peut même prendre en charge, de manière plus ou moins transparente, une partie des reformulations qui assurent un bon accès au contenu des documents indexés.

Néanmoins, lorsque Tudhope et d'autres montrent comment le thesaurus peut améliorer le fonctionnement du système documentaire, ils pointent aussi sur une de ses lacunes les plus importantes. Les relations entre termes apportées par le thesaurus ne sont en effet pas suffisantes pour obtenir une structuration transversale suffisamment fine du champ sémantique dont le vocabulaire est donné, ainsi que des index que l'on construit en utilisant ce vocabulaire. Par exemple, il peut être souhaitable d'exploiter des relations autres que les relations hiérarchiques, comme les relations méréologiques (entre tout et partie), les relations de localisation, de causalité, ou certaines autres plus spécifiques à un domaine applicatif.

Le premier problème est que les thesaurus peuvent introduire de telles relations entre les notions qu'ils contiennent, mais qu'ils ne le font que de manière implicite. Soergel dans [SLL$^+$04] rappelle par exemple que la relation méréologique est parfois exprimée aussi bien par l'intermédiaire des relations hiérarchiques que des relations d'association, et ce au sein d'un même thesaurus. Il faut noter que ce défaut peut avoir des répercussions très importantes : si un thesaurus associe une interprétation floue aux relations qu'il présente, et en particulier à la relation

[12]En recherche d'information, le rappel désigne le rapport entre le nombre de réponses correctes renvoyées effectivement par le système et le nombre de réponses correctes qu'il aurait dû renvoyer. Cette notion est complémentaire du silence.

[13]La précision désigne le rapport entre le nombre de réponses correctes et le nombre de réponses total renvoyées par le système. Cette notion est complémentaire du bruit.

hiérarchique, on peut douter qu'il suffise à apporter un contexte d'interprétation suffisamment précis pour garantir la continuité sémantique. Il serait donc nécessaire de distinguer clairement ce qui relève de l'organisation hiérarchique de la structuration horizontale, et d'introduire explicitement les types de relations qui correspondent aux associations entre notions d'un thesaurus. De telles approches existent, mais elles sont excessivement rares, les concepteurs refusant de complexifier inutilement les ressources, surtout si celles-ci sont systématiquement exploitées par des opérateurs humains, qui peuvent compter sur leurs connaissances afin de lever les ambiguïtés relationnelles.

Le second concerne la capacité à créer des relations de manière « dynamique » *au sein même des index* pour rendre compte de la manière dont le contenu d'un document peut associer les notions qu'il présente. Or ceci est impossible dans le cas d'un langage documentaire strictement thésaural n'autorisant pas de relation syntaxique[14] entre les descripteurs.

Comment rendre alors compte de la richesse du contenu documentaire audiovisuel, de son organisation comme de sa matière ? Dans l'une des applications du projet OPALES, nous avons par exemple un documentaire qui contient une séquence d'archives introduisant un discours historique, et une séquence d'interview vulgarisant un aspect technique. Si toutes ces notions (`documentaire`, `séquence d'archives`, `interview`, `aspect technique`, `discours historique`, `vulgarisation`) sont livrées pêle-mêle dans l'index du document, il est fort probable que l'interprétation du chercheur soit quelque peu différente de ce que l'indexeur a voulu communiquer : un documentaire contenant une séquence d'archives présentant une interview d'un chercheur sur l'histoire de la vulgarisation des techniques scientifiques aurait le même index.

C'est pourquoi l'indexation à l'INA utilise toujours, en complément du thesaurus qui norme le vocabulaire de certains champs, d'autres zones en texte libre (cf. tableau 1.4). On peut ainsi créer des expressions mettant en jeu des relations entre les notions reconnues dans un document. Mais, à nouveau, on retombe dans le piège de la variabilité textuelle. Il importe donc de créer un moyen d'exprimer de manière rationnelle, contrôlée, des *structures* rendant compte de la complexité du document à décrire.

Descripteurs	... maladie cardio vasculaire (athérosclérose) ; infarctus ; cholestérol ...
Descripteurs secondaires	... endoscopie ...
Résumé	... - Après une reconstitution d'un homme faisant un infarctus, Jean Charles FRUCHART explique à l'aide d'image endoscopique comment l'athérosclérose, maladie cardio-vasculaire par excès de cholestérol, est la première cause de mortalité en France ...

TAB. 1.4 Utilisation du texte libre en complément des mots-clefs

(extrait du tableau 1.1)

1.3.3 Structurer les index

En fait, dès le niveau de la notice, on doit se poser le problème de la structuration de l'information présentée dans la description du document, entre informations catalographiques,

[14]Dans le monde documentaire, on y fait ainsi référence car elles jouent pour les descripteurs et les descriptions un rôle semblable à celui de la syntaxe pour les mots de la langue et les phrases.

description thématique, mais aussi informations pour la gestion des supports physiques ainsi que de la notice elle-même en tant qu'élément du système documentaire. En ce qui concerne le cas du texte, l'organisation de la notice en un certain nombre de champs (voir section 1.1) permet d'obtenir une structuration plus efficace pour la recherche.

Mais ce problème de structuration de l'information se repose, à l'intérieur de la notice, pour les champs qui concernent la description. Le document audiovisuel est en lui-même un objet riche, mais de plus il peut être amené à rendre compte de situations ou de concepts complexes. Si une application requiert un niveau de description très fin, les informations à saisir dans l'index peuvent prendre une forme élaborée. Par la suite, nous allons nous attacher à un exemple tiré d'une de nos expérimentations, que nous présenterons plus complètement en section 5.2. Il s'agissait de décrire des vidéos relatives au domaine de la médecine – plus précisément de la chirurgie cardiaque – d'un point de vue attaché à l'analyse de la mise en forme audiovisuelle de concepts scientifiques, point de vue jouet, mais très proche dans ses préoccupations de l'un des cas d'utilisation du projet OPALES. En l'occurrence, une des vidéos peut être décrite comme une

> « émission qui contient une séquence où un professeur témoigne dans son bureau sur la maladie bleue [une maladie cardio-vasculaire], et où une animation explique ce qu'est une sténose, et comment le fonctionnement [anormal] du cœur est la cause de la maladie bleue ».

En supposant que l'on dispose dans un thesaurus des notions adéquates – de fait, de nombreux termes parmi ceux qui vont suivre peuvent être trouvés dans le thesaurus de l'INA – l'index créé contiendra les notions `émission`, `séquence`, `témoignage`, `professeur`, `bureau`, `maladie bleue`, `animation`, `sténose`, `fonctionnement`, `coeur`. Un tel index, non structuré, pose, on vient de le voir grâce à un exemple similaire, des problèmes d'interprétation : en exagérant à peine, on peut imaginer qu'un tel index pourra répondre à une requête sur le fonctionnement des bureaux. De plus, ce qui est particulièrement important dans une application traitant de documents audiovisuels, on ne distingue pas les procédés de construction ou de mise en forme audiovisuelle des thèmes que ces documents abordent, ou des personnes, des lieux et des actions qu'ils montrent.

Afin de remédier à cela, on peut structurer à leur tour les champs d'indexation en sous-champs plus spécifiques, chacun rendant compte d'un aspect du document audiovisuel. Ceci implique de construire une grille de description, un *formulaire*, qui sera utilisée pour chaque index.

Cette grille peut être inspirée de grilles d'analyse standard documentaire : un champ pour ce qui se passe (« Quoi ? »), un champ pour les personnes qui interviennent (« Qui ? »), un pour les lieux (« Où ? »), etc. A l'INA, dans le cadre d'une exploitation dite « thématisée », où des corpus de documents ou séquences sont constitués afin de traiter d'un sujet, d'un événement ou d'une personne particulière, on a essayé de complexifier la structure des index traditionnels, en introduisant des *types de descripteurs*, qui en étant accolés à des descripteurs dont il précisent la catégorie permettent en quelque sorte de structurer les champs d'index en sous-champs :

- DET pour les thèmes du segment documentaire ;
- DEL pour les lieux où se déroule l'action ;
- DEI pour des descripteurs relatifs à la construction des images ;
- DSO pour les descripteurs liés à la bande sonore.

Ainsi, notre exemple pourrait être décrit par les descripteurs présentés dans le tableau 1.5.

On peut par conséquent commencer à raffiner les requêtes en se focalisant sur un aspect particulier, tout en étant sûr que les descriptions, contraintes par le formulaire, seront correcte-

DET	`maladie bleue`
DET	`sténose`
DET	`cœur`
DEL	`bureau`
DEI	`animation`
DSO	`interview`

TAB. 1.5 Structuration d'un ensemble de descripteurs

ment utilisées par le système[15]. On a donc une structuration qui d'une certaine manière aide à maintenir la continuité interprétative entre l'indexeur et le chercheur. On peut noter que cette approche garantit également une certaine cohérence quant au traitement documentaire : tout comme le thesaurus contraint le vocabulaire de la notice, le formulaire permet l'instauration d'une politique de contrôle éditorial des index.

Cette approche peut cependant sembler trop rigide pour tous nos besoins. Elle ne permet de rendre compte que de relations figées lors de la définition du formulaire, ce qui posera problème lorsqu'on voudra créer un index légèrement différent. De plus, le niveau d'application de ces relations reste implicite : si l'un de leurs relata est connu (la valeur d'un champ), l'autre n'est pas immédiatement accessible. La notion référencée dans un champ joue-t-elle le rôle indiqué par rapport à l'entité décrite par le formulaire – en principe, le segment documentaire – ou par rapport à la valeur d'un autre champ du formulaire ? Et si oui, lequel ? De l'index pseudo-structuré de notre exemple, on peut déduire que c'est l'émission décrite qui parle de la sténose, et non le professeur qui y intervient. A moins d'introduire des structures de type « sous-formulaires », qui pour chaque entité présente dans les champs offriraient à nouveau la possibilité d'une description structurée. Ce genre de mécanisme peut être utile, mais il reste rigide, et pour l'instant n'apparaît pas parmi les fonctionnalités couramment offertes. Un formulaire ne peut tout anticiper et rester facilement utilisable.

Certains thesaurus proposent des mécanismes permettant de préciser le rôle, dans une situation décrite, des entités auxquelles réfèrent les descripteurs : les *facettes* [Map95]. A l'intérieur d'un même champ sémantique, on va classer les notions – la relation hiérarchique est alors de type genre-espèce – suivant des catégories de haut niveau qui indiquent implicitement des relations entre ces notions. L'exemple du tableau 1.6, extrait du thesaurus de l'INA, montre comment, dans une même branche hiérarchique, les notions spécialisées peuvent être envisagées selon le rôle qu'elles joueront, ici dans un schéma actanciel générique : des *personnes* participent à des *actions* pratiquées selon des *techniques* données, dans des *lieux* spécifiques, etc.

Mais ces liens, comme ceux donnés par un formulaire, sont figés, et restent formulés à un niveau souvent trop général par rapport aux besoins applicatifs : que faire si on veut expliquer, dans le cadre du schéma implicite de notre exemple, qu'une notion, en plus de dénoter une personne, sera associée au rôle de *patient* ou d'*agent* d'une action ? Il faudrait introduire une autre hiérarchie correspondant à ces rôles, en émettant l'hypothèse forte – et fausse – qu'une personne ne peut pas être agent dans un cas et patient dans un autre. Finalement, ces rôles étant intégrés « en dur » aux mots-clefs, il est difficile de les faire bénéficier d'un traitement particulier, conforme aux besoins de l'application envisagée, à moins d'adapter le système documentaire lui-même au thesaurus employé, ce qui lui ferait incontestablement perdre en souplesse et généricité.

[15]En pratique, dans un système documentaire informatisé, les notices structurées en champs pourront être en effet stockées dans des bases de données, ce qui permettra une interrogation et un accès immédiats aux données de chaque champ, des croisements entre valeurs de différents champs, etc.

```
théâtre
       $infrastructure-théâtre
              théâtre de plein air
              UP :théâtre de verdure
              théâtre-bâtiment
       $matériel-théâtre
              masque de théâtre
              rideau de théâtre
       $organisme-théâtre
              troupe théâtrale
peinture-art
       $matériel-peinture-art
              palette de peintre
              UP :palette
              pinceau
       $personne-peinture-art
              artiste peintre
              modèle
       $technique-peinture-art
              peinture murale
                     fresque
                     peinture rupestre
              peinture sur tissu
```

Tab. 1.6 Les facettes dans le thesaurus de l'INA

Un $ indique une notion qui ne peut être employée comme descripteur

Pour ces raisons – et d'autres, voir [Blo04] – les systèmes utilisant des thesauri à facettes sont souvent considérés comme trop lourds à employer, aux regards des améliorations qu'ils apportent concrètement.

A l'INA, afin d'exprimer explicitement des relations entre descripteurs, on utilise parfois des *précisions d'indexation* [Pic96]. Il s'agit au sein d'un index d'établir un lien entre deux termes. Ces précisions sont le plus souvent utilisées pour expliciter le sens d'un mot outil trop général sinon (`vérification(contrôle technique)`), pour indiquer, après le descripteur adapté, le nom propre d'un lieu, d'une entreprise ou d'une institution (`musée(Orsay)`), d'une manifestation artisitique ou non (`pièce de théâtre(Andromaque)`), ou d'une autre entité (`logiciel(Word)`).

De cette façon, on peut créer au niveau des index des relations entre les composants de ces index, semblables aux relations que l'on pouvait trouver dans le thesaurus : ainsi la relation d'instanciation[16] entre un descripteur et un nom propre se rapproche de la relation hiérarchique. Néanmoins, le problème qui se posait dans le cas du thesaurus réapparaît ici. On met bien en relation deux descripteurs, mais on ne peut pas utiliser de notion spécialisée, adaptée à une application, pour décrire cette relation.

Plus intéressante est la possibilité de se servir d'une précision d'indexation pour exprimer un lien entre deux ou plusieurs entités, comme dans `accord (Israël Palestine)` ou bien `ligne-transport (Paris Lyon)`. Dans ce cas, le terme à valeur relationnelle est bien issu du thesaurus, et peut être interprêté de manière conforme à ce qui est attendu. Il semble ainsi faisable d'exprimer les liens que l'on voulait faire figurer dans notre index : par exemple `témoignage(professeur)` et `explication (animation sténose)` semblent faire partie des rapports autorisés par les précisions d'indexation.

Cependant, l'utilisation de telles précisions se fait de manière *ad hoc* : en l'absence de consigne claire à ce sujet, tous les termes peuvent prendre une connotation relationnelle. Et les créateurs des index n'ont pas de moyen de préciser le statut des éléments liés dans la relation, ce qui peut être gênant si on veut expliciter une relation orientée. Après tout, on peut interprêter `vérification(contrôle technique)` autant comme la vérification d'un véhicule que constitue le contrôle technique, que comme la validation des procédures et des moyens employés lors d'un contrôle technique.

De fait, l'utilisation de ces relations n'est pas vraiment contrôlée : même si le vocabulaire, et notamment les termes relationnels, bénéficient de la normalisation apportée par un thesaurus, il n'y a pas de contrainte sur leur utilisation dans une précision d'index. Si on veut rester cohérent avec l'interprétation de départ des précisions apportées à `témoignage` – la précision donne la personne qui témoigne – on ne devrait pas pouvoir rencontrer dans un index l'expression `témoignage(sténose)` indiquant qu'un témoignage a pour objet une sténose.

Le risque est finalement de retomber dans les travers de l'utilisation du langage naturel, la richesse expressive en moins. C'est pourquoi ces précisions d'indexation sont assez rarement rencontrées, les documentalistes préférant utiliser les champs textuels classiques.

La structuration des index est donc encore un problème pratiquement ouvert. Les solutions de type formulaires apportent suffisamment de contrôle et d'expressivité pour pouvoir être employées dans les approches peu exigeantes, mais elles sont trop rigides pour pouvoir être adaptées à toutes les applications, surtout celles qui requièrent des descriptions d'objets aussi complexes que les documents audiovisuels. L'impératif de structurer des relations transversales entre descripteurs, en particulier, est une difficulté majeure pour ces approches. De l'autre côté, une approche de

[16]Relation entre une catégorie et un individu classifié dans cette catégorie.

mise en relation *ad hoc* de descripteurs à l'intérieur des index n'est pas facilement adoptable sans perdre la continuité interprétative que nous souhaitions. Il semble qu'un paradigme pertinent d'indexation relationnelle soit toujours à introduire dans l'approche documentaire.

1.4 Conclusion

L'enjeu majeur des systèmes documentaires est de rendre possible un accès raisonné au contenu des documents, adéquat aux applications envisagées. Ceci est encore plus important dans le cas de l'audiovisuel, forme qui ne prescrit pas naturellement d'interprétation qui permettrait de manipuler aisément ce contenu : on doit alors impérativement effectuer une *indexation*. Au cours de ce processus, une des nombreuses significations possibles du document audiovisuel, celle qui est pertinente par rapport à l'application envisagée, sera explicitée, puis reformulée de sorte à être exploitable dans le cadre d'une application particulière. Traditionnellement, les systèmes d'indexation utilisent la forme textuelle, qui a l'avantage sur l'audiovisuel de s'inscrire dans un système fonctionnel d'assignation de sens.

Cependant, nous avons rappelé qu'une exploitation correcte ne peut être assurée que si une réelle *continuité sémantique* s'applique à l'ensemble de la chaîne de description et de recherche. Dans le cas précis de l'index, il faut que la forme documentaire retenue soit à même de garantir une compréhension précise et univoque : dans un système documentaire, il faut obtenir des descriptions dont l'interprétation est contrôlée.

Le premier des moyens est de contrôler l'accès aux descriptions. A l'INA, quand on le peut, on fait en sorte que les documentalistes puissent assister ou effectuer eux-mêmes les recherches parmi les descriptions qu'ils ont créées : on pallie ainsi une partie des variations observées dans les pratiques d'explicitation et de reformulation.

Néanmoins il est évident qu'une telle approche ne peut être retenue pour toutes les applications. Des techniques existent, qui en agissant sur la manière dont on exprime les index permettent un contrôle de l'interprétation et une exploitation moins exigeante en investissement humain. Nous avons vu qu'il est courant dans le monde documentaire de se tourner vers des langages s'appuyant sur des thesauri, qui normalisent les interprétations des notions utilisées dans les descriptions. On montre cependant que parfois ces thesauri sont trop limités, spécialement par leur manque de relations sémantiques dédiées à des cas applicatifs précis. En effet, en particulier dans le domaine de la description audiovisuelle, il est nécessaire de pouvoir exprimer des relations entre les composants d'un index. Et les techniques traditionnelles ne suffisent pas : on peut normaliser la structure d'une description, et la manière dont cette structure est interprétée, mais l'obtention d'un contrôle satisfaisant se fait souvent aux dépens de la finesse et de la richesse des expressions autorisées.

On retrouve là un problème général aux solutions documentaires utilisées couramment : des progrès énormes ont été accomplis au regard des performances des systèmes de gestion et d'accès aux descriptions, mais en général on peine à obtenir un compromis satisfaisant entre la possibilité de créer des indexations riches et celle de les utiliser de manière cohérente et efficace. A l'heure où la numérisation et la mise en réseau des contenus s'accélère, et que par conséquent l'utilisateur qui accède aux descriptions dudit contenu est de plus en plus éloigné[17] de leur créateur, ce problème

[17] Cet éloignement est à la fois physique, puisque l'on peut accéder à des bases documentaires distantes du lieu d'où on les consulte, et conceptuel, puisque l'on n'est plus nécessairement en contact intellectuel avec le contexte interprétatif qui motive la création des descriptions. A ces deux aspects s'ajouter un troisième, celui de l'éloignement « technique » : le schéma classique utilisateur/système de données devient plus compliqué, un système pouvant à son tour faire appel à d'autres systèmes afin de bénéficier de leurs services. Pour cela, il est nécessaire d'obtenir des conditions d'interopérabilité – entre systèmes – et de compréhension – entre utilisateurs

est de plus en plus préoccupant.

Il convient alors de se tourner vers l'informatique, science de la manipulation de l'information, pour proposer des solutions techniques susceptibles d'aider le déroulement de la description et de la recherche documentaire. Concrètement, nous allons chercher un support de description qui sera exploitable de manière pertinente par un système d'assistance à l'accès au contenu. Ici, nous n'allons pas suggérer de mécanisme physique particulier d'encodage de l'information, mais plutôt nous préoccuper de la création de langages de description des contenus documentaires, en nous focalisant sur le niveau conceptuel de telles descriptions. En effet, plus que sur la représentation générale des contenus des documents, nous nous concentrons sur la description des *significations* que peuvent prendre ces contenus pour des applications spécifiques[18].

Au cours de ce chapitre, nous avons montré que les langages documentaires actuels présentent des lacunes au regard de la description des interprétations documentaires. Pour obtenir un système d'indexation et de recherche plus performant, il est important d'apporter des améliorations aux langages existants, en proposant des dispositifs de création de langages et d'index qui puissent être évalués positivement par rapport aux fonctionnalités résumées ici.

F1 : Expressivité. Il est indispensable de pouvoir produire au sein du système d'indexation des descriptions suffisamment riches et précises de la signification des documents, afin de les restituer le plus fidèlement possible. Ceci implique de traiter correctement les points de la forme et de la substance des descriptions.

- *forme de la description* : nous cherchons à adopter un paradigme relationnel de représentation des interprétations, qui puisse rendre compte d'un contenu composé de notions associées les unes aux autres. Une description doit contenir les concepts issus de l'interprétation des contenus, mais aussi les relations qui les articulent en une signification qui est nécessairement plus complexe qu'une simple liste de notions.
- *substance de la description* : il s'agit pour chaque application d'un système d'indexation de pouvoir définir un vocabulaire de description, à la fois pour les concepts présentés et les liens les unissant. Les pratiques visées par les indexations emploient des notions dont le sens spécifique doit pouvoir être défini explicitement ; l'emploi de ces notions lors de la création des descriptions est une condition nécessaire de pertinence pour le système.

F2 : Contrôle. La continuité sémantique exige, on l'a vu, le recours à des mécanismes de contrôle des interprétations, que ce soit lors de leur formulation, ou lors de leur accès et de leur compréhension, lorsqu'un utilisateur ou le système lui-même doit opérer une forme de ré-interprétation nécessaire à leur exploitation. Là encore, on peut décomposer la tâche suivant les aspects précédemment évoqués :

humains et systèmes – qui complexifient d'autant le problème de la continuité sémantique.

[18]Cette vision nous distingue quelque peu de celles dictant l'élaboration des langages documentaires *stricto sensu*, qui doivent effectivement s'intéresser à la gestion des aspects logiques de l'organisation du contenu des documents. De telles approches ayant fait l'objet de réflexions approfondies à l'INA, notamment dans le travail de thèse de Raphaël Troncy [Tro04], il ne nous appartient pas de proposer ici nos propres solutions. Par rapport à ce qui a été évoqué à la page 19, nous n'abordons donc pas directement les étapes de localisation et de structuration. Pourtant il apparaîtra au lecteur des chapitres suivants que nous ne nous refusons pas à décrire certains aspects de la structuration logique du contenu : nous avons d'ailleurs travaillé autour de certaines propositions de [Tro04]. Il s'agit juste de reconnaître que dans notre optique, la réponse à ce besoin est subordonnée à sa reconnaissance dans une application précise. Même dans le cadre audiovisuel, la signification « utile » d'un document peut dans certains cas faire abstraction de son organisation.

– *substance de la description* : les descriptions doivent effectivement être créées à l'aide du vocabulaire spécialisé évoqué ci-dessus. Il faudra faire en sorte que l'accès aux notions mobilisées dans les descriptions puisse s'effectuer en ayant accès à la signification qui a été arrêtée lors de la définition de ce vocabulaire, par la construction d'un contexte d'interprétation tel que le réseau proposé par un thesaurus.
– *forme de la description* : dans la perspective de descriptions orientées par des applications précises, il est naturel d'envisager la mise en place de processus de contrôle de la forme finale de ces descriptions. Chaque application vient avec des exigences de cohérence : si on veut obtenir des expressions descriptives pertinentes, il faudra le plus souvent guider, voire restreindre la manière dont les termes du vocabulaire sont articulés les uns avec les autres. En somme, il s'agit de mettre en place pour nos index ce que l'on peut assimiler à un contrôle éditorial, qui limitera la variabilité[19] des index produits.

F3 : Manipulation des descriptions. Dans le cas des systèmes documentaires traditionnels, l'intérêt du contrôle sémantique réside aussi dans les fonctions d'assistance que les procédés d'indexation rationnelle autorisent. Dès lors qu'une description voit son sens précisé, normé par un contexte d'interprétation et d'exploitation, on peut mettre en place des traitements automatiques qui prennent en charge une partie des tâches élémentaires qui incombent traditionnellement au documentaliste. Dans un système thésaural, on peut utiliser les relations sémantiques entre notions pour reformuler les requêtes. Dans un cadre encore plus précis, où les index seraient explicitement structurés et comprendraient des notions définies clairement lors de la création du langage d'indexation, des mécanismes d'assistance automatique plus efficaces, s'appuyant par exemple sur des *raisonnements* conduits par le système lui-même, sont envisageables.

Lorsqu'on observe des exemples de reformulations de requêtes documentaires, tel que celui du tableau 1.2, on voit que le paradigme de description que nous recherchons doit permettre la manipulation de symboles munis du statut de véritables *connaissances*. Dans la série de reformulations qui est présentée, on se repose sur le fait que les différents types de supermarchés constituent des notions qui sont proches, et que si on substitue l'une à l'autre, on modifie assez peu l'interprétation de la description. Ou encore qu'il est utile d'exploiter pour une recherche documentaire la connaissance selon laquelle une entité – un magasin – est liée à une autre – une ville. Il faut donc chercher à reproduire les procédés d'association qui sont utilisés afin de reformuler les requêtes, et qui reposent sur les connaissances qu'ont les opérateurs humains du domaine d'application du système documentaire. Au-delà du rôle de clarification terminologique classique des langages documentaires, les structures présentées doivent être suffisamment précises pour pouvoir être exploitées par des systèmes de raisonnements formels, capables d'assister les documentalistes ou utilisateurs non spécialistes de la recherche documentaire dans leur travail.

Pour l'instant, il est encore bien sûr déraisonnable d'espérer un système générique capable d'exploiter une base documentaire de manière aussi fine qu'un humain. Mais on peut tout à fait envisager la spécification d'outils d'assistance en fonction des besoins d'une pratique ou d'un domaine particulier, afin de leur faire effectuer quelques raisonnement basiques, au moment même

[19]On peut remarquer que cette variabilité, revêt deux aspects, tout aussi problématiques :
– une variabilité inter-indexeur : deux indexeurs différents peuvent délivrer des formulations différentes pour des contenus identiques ;
– une variabilité intra-indexeur : un même indexeur à deux moments peut conduire des stratégies d'interprétation ou de formulation différentes.

où l'accroissement des volumes de données rend l'application « manuelle » de ces raisonnements plus difficile.

D'un point de vue architectural, une telle évolution doit être correctement répercutée dans les trois composants logiques que l'on peut isoler dans un système de description documentaire :

- **C1 : Langage de description.** On doit pouvoir spécifier pour les descriptions un vocabulaire de concepts et de relations qui définit ses composantes à l'aide de connaissances exploitables en vue d'une application donnée. Il s'agit tant de connaissances définitoires que de règles de cohérence contrôlant l'emploi du vocabulaire.
- **C2 : Descriptions.** L'ensemble des descriptions doit être considéré comme un ensemble de connaissances interprétables tant par le système que l'utilisateur final. La création des descriptions résulte d'un engagement de la part de l'indexeur par rapport au référentiel de signification défini par le langage ; il importe de saisir et d'exploiter ces données en fonction de cet engagement.
- **C3 : Système.** L'exploitation de descriptions et de langages de descriptions élaborés nécessite évidemment le recours à un système autorisant la création et la manipulation des unes en fonction des autres. Idéalement, un tel système doit être générique, capable d'adapter son comportement en fonction des connaissances dont il dispose : les descriptions, bien sûr, mais aussi les langages, qui varient en fonction des pratiques considérées.

Nous allons voir à présent qu'il est naturel dans un tel cadre de se tourner vers des *systèmes à base de connaissances*. Issus des recherches en intelligence artificielle, certains travaux du domaine de la représentation des connaissances [Kay97] fournissent en effet des outils permettant la création de ressources riches et mobilisables au sein de processus d'inférence. L'interprétation du langage documentaire et des descriptions construites va être augmentée d'une interprétation formelle exprimant la manière dont un système peut exploiter ces éléments de connaissance. L'enjeu devient alors de choisir un formalisme de représentation pertinent, et de spécifier les ressources conceptuelles – en définissant le langage de description – que le système pourra manipuler. Cette dernière tâche est l'objet de l'*ingénierie des connaissances* [CZKB00], qui, pour concilier le souci de modélisation formelle et celui de l'intelligibilité des ressources, condition nécessaire à l'obtention de réelles connaissances, dépasse la notion de thesaurus et propose celle d'*ontologie*.

Chapitre 2

Ontologies et systèmes à base de connaissances pour la description conceptuelle de documents audiovisuels

2.1 Introduction : représenter au niveau de la connaissance pour indexer

Nous cherchons à représenter différemment nos index, afin que leur création et leur manipulation tiennent mieux compte du contexte dans lequel ils sont produits et exploités. Les utilisateurs dans un projet comme OPALES ont en effet besoin d'une indexation très fine, à même d'exprimer les relations entre les différents éléments du contenu des documents audiovisuel, que ces éléments soient d'essence documentaire ou thématique. Ceci impose de pouvoir utiliser un langage d'indexation élaboré, dont le vocabulaire est adaptable en fonction du point de vue applicatif considéré.

Pour faciliter l'accès aux documents, ce langage doit également permettre des traitements qui utilisent la matière des index, et ce d'une façon également adaptée au point de vue applicatif. A l'heure où les besoins documentaires deviennent plus complexes, l'exploitation de l'index, point de passage obligé entre des acteurs – documentaliste, expert du domaine, utilisateur non-expert – de plus en plus isolés face au système documentaire, est cruciale. Il faut donc voir comment introduire des mécanismes plus fins de représentation des interprétations dans le système documentaire, et comment ces mécanismes peuvent être employés avec des connaissances qui spécifient le comportement du système dans un contexte applicatif particulier.

Nos besoins, on l'a vu, concernent à la fois des fonctions liées :

1. à l'expressivité des descriptions : nous voulons des index complexes, dont les éléments peuvent être articulés par des liens typés spécifiques à l'application envisagée ;
2. au contrôle de l'expression : le système doit fonctionner de manière à garantir une continuité d'interprétation entre les utilisateurs divers, qu'ils soient créateurs d'index ou chercheurs. Ceci implique que les index doivent être créés et manipulés conformément à un référentiel partagé de signification ;
3. à la manipulation des descriptions pour l'assistance à la recherche d'index : le système doit être capable de mettre en œuvre des mécanismes pour rapprocher les index des requêtes, à chaque fois que cela est possible. Ceci implique de pouvoir spécifier un maximum de connaissances de raisonnement pertinentes pour le domaine ciblé par l'application.

Dans un cadre informatique, l'intelligence artificielle apporte des techniques de *représentation des connaissances* (RC) permettant la création de langages formels appliqués à des domaines d'application précis. Articulés à des processus d'*acquisition des connaissances* cohérents [ALR96], ils permettent d'obtenir des systèmes à base de connaissances (SBC) intelligibles dont le comportement, on va le voir, répond aux fonctionnalités que nous avons dégagées.

Depuis assez longtemps, une partie des travaux de l'intelligence artificielle, reprenant les termes introduits par Newell dans [New82], repose en effet sur la distinction entre le niveau *symbolique*, qui est celui de l'implémentation formelle des programmes, et celui des *connaissances* manipulées par des agents rationnels. Selon ces approches – on se reportera à [SAA+99] pour un exemple complet – il est possible et souhaitable de caractériser les systèmes (comportements observables, entités représentées) en termes des connaissances manipulées par les experts de leur domaine d'application. Le vocabulaire, les mécanismes de traitement des données, etc. sont définis indépendamment de leur encodage dans un programme. On espère ainsi, en rapprochant les entités représentées et les traitements effectués par le système de la compréhension que peut avoir un agent expert du domaine visé, améliorer le processus de conception et d'utilisation des outils informatiques.

Comment de tels mécanismes peuvent-ils être mis en œuvre ? Dans le monde « réel », nous avons des *cas* correspondant aux situations diverses rencontrées dans une application, cas qui sont décrits à l'aide de la langue naturelle, qui permet d'en stabiliser une représentation intelligible. Par exemple, nous pouvons décrire une séquence d'entretien avec un médecin comme étant l'interview d'un expert si les notions d'interview et d'expert sont celles qui nous intéressent en priorité.

Dans un système documentaire, on a vu qu'il fallait plus de rigueur dans l'expression des index. Il faut réaliser un effort d'abstraction pour isoler dans un cas les éléments pertinents et la manière dont ils sont organisés, leur configuration. Concevoir une telle abstraction présuppose une représentation théorique du domaine que l'on ne peut qu'approcher, et cela au prix d'un effort conséquent d'acquisition des connaissances relatives à l'application.

Il faut en particulier isoler et organiser les concepts[1] qui permettent d'organiser les connaissances rencontrées dans le domaine. Un tel travail de conceptualisation est nécessaire, si l'on veut créer des expressions manipulables par un programme informatique, conçu pour faire des calculs à partir de ce qui ne sont que des formes : toute expression informatique, même si elle est élaborée, repose finalement sur un codage binaire. Dans notre situation, les index devront être représentés de sorte à avoir une signification formalisée. De la signification intelligible des éléments composant les index, celle qui résulte de l'interprétation humaine, ancrée dans la verbalisation et la pratique d'un domaine applicatif, on va abstraire des spécifications qui s'inscrivent dans un cadre interprétatif formel [Bac96]. La signification des expressions s'obtient alors à partir des significations de leurs composants élémentaires, processus qui s'appuie traditionnellement sur des interprétations en termes d'ensembles et d'opérations ensemblistes. Cette signification formelle sera alors facilement exploitable pour obtenir une spécification des inférences autorisées pour les programmes de raisonnement – ces inférences constituent une interprétation calculatoire

[1]Le terme *concept* est ambigu, du fait de son emploi – suivant des acceptions différentes – dans diverses disciplines, son sens variant même parfois à l'intérieur d'un même champ. En représentation des connaissances, on verra par exemple qu'un « concept » pour les logiques de description correspond à un « type de concept » pour les graphes conceptuels. En première approche, on peut suivre l'acception simpliste qui associe le terme *concept* à tout ce qui peut être pensé. Pour une discussion beaucoup plus intéressante sur la signification de cette notion, le lecteur est invité à se tourner vers [Bac04a], pp.128-140.

des index, elles explicitent leur signification pour la machine qui les traitera : on parle alors de signification computationnelle.

Dans un système à base de connaissances, les expressions sont contrôlées par le recours à un langage de représentation. Celui-ci fournit le vocabulaire – les *primitives*[2] – et la grammaire pour créer des expressions valides, ainsi que les éléments permettant de leur associer une interprétation formelle puis opérationnelle. Il faut alors construire une spécification de ce langage qui soit correcte vis-à-vis de la conceptualisation. Pour cela, on utilise des artefacts élaborés, les *ontologies* (nous en discuterons plus en détail dans la section 2.2.2), dont l'objectif est d'expliciter autant que faire se peut le sens du vocabulaire proposé par cette conceptualisation.

Adosser un langage de représentation de connaissances à une telle ontologie – on parle d'*engagement ontologique* [GCG94] – permet de donner une signification aux index qui seront construits par son moyen, aussi bien pour le système chargé de leur manipulation que pour les utilisateurs humains qui vont les créer et les visualiser. Le schéma de la figure 2.1 donne un aperçu des rapports entre les différents aspects constituant l'environnement et les éléments des SBC.

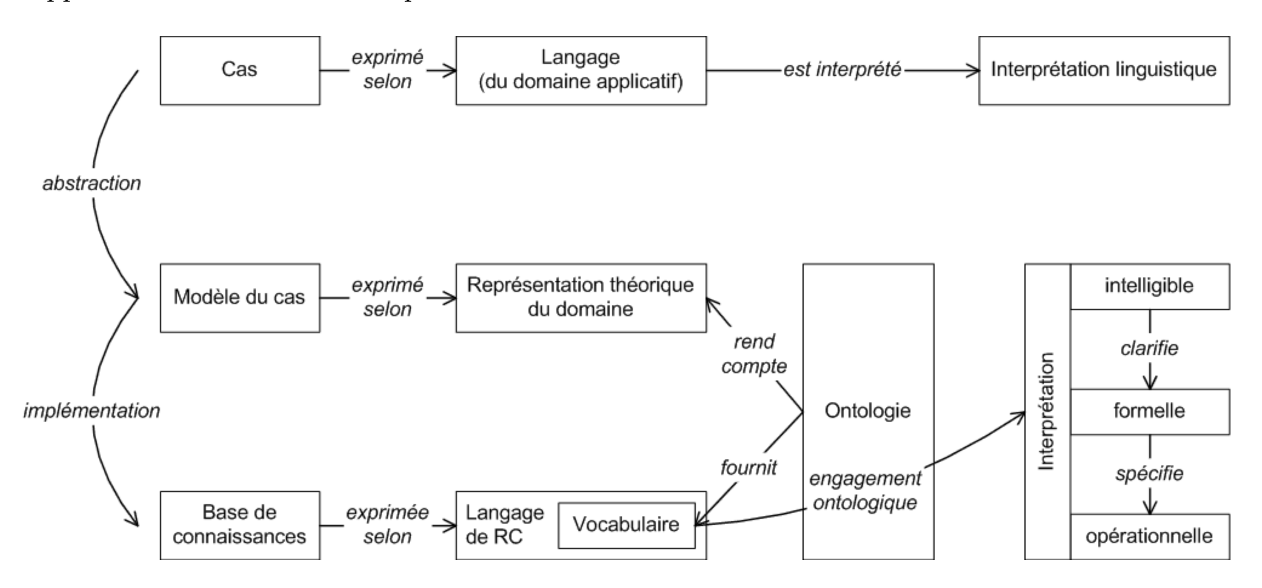


FIG. 2.1 Les ontologies au sein des SBC.

On va voir que dans ce cadre il devient naturel d'exprimer des connaissances complexes, plus proches des besoins de l'indexeur que les index construits avec les systèmes documentaires classiques. En particulier, il est possible de choisir un paradigme relationnel pour la description : les index seront déclinés au niveau de la connaissance en des réseaux d'instances de *concepts* structurés par des *relations*.

On vient d'évoquer que pour les SBC il était nécessaire de recourir à des ontologies pour normer le vocabulaire utilisé, et lui donner une signification précise, exploitable par l'ensemble

[2]Ce sont les symboles utilisés par les langages de RC qui préexistent à l'expression des connaissances dans une base de connaissances. Nous considérons en effet ici qu'un langage de RC inclut tout ce qui est nécessaire à la représentation de faits représentant des situations particulières. Cette acception du terme « langage » est différente de nombreuses approches, pour lesquelles les langages de RC renvoient plutôt à des méta-langages, dont la vocation est celle de créer le vocabulaire permettant de représenter les faits. De tels langages ne contiennent que du vocabulaire d'essence logique – les constructeurs syntaxiques autorisant la composition d'expressions – et épistémique – précisant les différents types de connaissances dont relèvent les éléments de vocabulaire introduits.

des acteurs rencontrés, logiciels ou humains. Nous définirons plus précisément ces objets, et détaillerons leur rôle vis-à-vis des langages de représentation. Nous allons voir que les ontologies instaurent des contraintes interprétatives propices à l'établissement de la continuité sémantique que nous recherchons, et que les spécifications formelles qu'elles apportent contribuent à optimiser la manipulation des index par le système, en généralisant le recours à l'inférence sémantique. Tout au long de cette démonstration, nous ne manquerons pas de comparer les fonctionnalités de systèmes utilisant ces ontologies avec celles des systèmes exploitant les objets *a priori* proches que sont les thesauri. Et nous illustrerons notre propos d'exemples tirés en particulier d'OPALES, qui utilise le formalisme de représentation et de raisonnement des graphes conceptuels.

Nous survolerons ensuite la mise en œuvre de ces principes dans des systèmes expérimentaux, depuis les précurseurs en matière d'annotation sémantique à ceux qui se concentrent sur le document audiovisuel. Nous verrons alors quels sont les obstacles à lever si on veut passer du stade de solutions de recherche « pure », complexes à mettre en œuvre, à celui de systèmes prenant mieux en compte les besoins rencontrés dans les usages réels.

2.2 IC et expressivité descriptive

2.2.1 IA et langages structurés

La représentation des connaissances et des raisonnements humains, depuis les travaux fondateurs inspirés par les mathématiques, comme pour [Fre71], la sémiotique – notamment par les travaux de Peirce – ou la psychologie cognitive [Qui68], s'oriente vers la création de *formalismes de représentation* élaborés. En n'allant pas tout de même jusqu'à la complexité offerte par la langue, les tenants de la représentation de faits et de concepts à un niveau formel proposent des paradigmes expressifs riches, où les entités abstraites de l'observation du monde sont liées les unes aux autres, ces liens pouvant très souvent être considérés comme des entités de premier plan, au même titre que les objets qu'ils relient.

Ainsi, la logique (du premier ordre) introduit la notion unificatrice de *prédicat*. Ces prédicats, s'appliquant aux constantes d'un domaine d'interprétation ou aux variables, peuvent être d'arité unaire ou multiple. Ils représentent les notions abstraites qui servent à déterminer les objets concrets des configurations du monde représentées, ainsi que leurs interactions. Dans le cas de notre application audiovisuelle, nous pourrons trouver des prédicats unaires comme `Emission`, `Interview` ou `Professeur`, mais aussi des prédicats binaires comme `Contient` ou `Participe`. Il sera ainsi possible de créer un ensemble de formules logiques (figure 2.2) rendant – partiellement – compte de l'exemple de la page 29 : l'émission indexée contient une interview – qui est une séquence – dans laquelle un des participants, un professeur, se livre à une explication d'une maladie cardio-vasculaire, la maladie bleue.

$$
\begin{array}{l}
Emission(emission_decrite) \\
MaladieCardioVasculaire(maladie_bleue) \\
\exists i, p\ Interview(i) \wedge Professeur(p) \\
\quad \wedge Contient(emission_decrite, i) \wedge Participe(p, i) \wedge Explique(p, maladie_bleue) \\
\forall x\ Interview(x) \rightarrow Sequence(x)
\end{array}
$$

FIG. 2.2 – Formules de la logique du 1er ordre représentant un extrait d'index

En regardant cet exemple, on sait que l'objet que l'on décrit, `emission_decrite`, peut être compris comme étant une émission, et contient un objet non nommé qui est une interview. Il

est important de noter que cette relation d'inclusion entre l'émission et l'interview est désormais formulée de manière explicite, et, comme le reste des assertions, exploitable dans un système de raisonnement, ou par un humain initié aux règles d'interprétation des constructions parfois complexes que constituent ces formules. Les prédicats unaires sont en effet interprétés formellement comme des ensembles, dont les variables et constantes qui les instancient dénotent des éléments.

Ce cadre purement logique apporte donc une rigueur formelle intéressante : on sait à présent que l'on se place, en employant de telles formules, dans un domaine d'interprétation, un ensemble d'individus régis par des contraintes qui dérivent de la signification des constructeurs du langage. Cette approche constitue un cadre représentationnel bien plus riche que celui des langages documentaires classiques : avec un thesaurus, il n'est pas envisageable de se référer explicitement à des relations entre individus. On en reste au stade de la composition des notions, composition dont le sens même n'est pas toujours spécifié[3].

Dans la perspective de la conception et de l'exploitation de systèmes à base de connaissances, cette logique n'est cependant pas jugée satisfaisante. Les processus de démonstration ne sont pas décidables, et les performances des démonstrateurs souvent jugées trop faibles. De plus, il s'agit d'un formalisme insuffisant d'un point de vue épistémologique : s'il permet l'encodage formel de nombreuses conceptualisations, il ne propose pas de cadre qui préconise telle ou telle organisation *a priori* pour celles-ci. De fait, il y a souvent une distance trop grande entre la connaissance, ressource structurée et structurante relevant d'un domaine d'application, et les interprétations ensemblistes sur lesquelles repose le langage logique.

C'est pourquoi des approches ont été proposées, qui, s'éloignant plus ou moins de ce mécanisme de représentation purement formelle, ont proposé de remettre la notion – certes floue – de connaissance au cœur de l'activité de conception des systèmes d'information. Si l'on veut que de tels systèmes fonctionnent de façon effective, il faut qu'ils puissent rendre compte d'un engagement plus complet par rapport à l'organisation des données qu'ils manipulent, et de la manière dont ces connaissances informatisées sont rattachées aux connaissances effectives du domaine d'application considéré.

Certaines démarches s'appuient sur des bases non-logiques. C'est le cas des *frames* proposées par Minsky [Min81]. Cette approche met en avant l'importance, au niveau psychologique, de structures récurrentes complexes et préconise la modélisation de ces structures – les *frames* – comme l'objectif principal d'une approche de représentation de connaissances. On modélise ainsi la connaissance sous forme d'objets complexes, renvoyant à des entités du domaine dont la caractérisation s'opère par l'emploi d'attributs – *slots* – dont les valeurs sont contraintes par des propriétés – *facets*. On pourra ainsi dire que l'on représente une *émission* comme une frame, qui possède un attribut *contient*, qui prendra ses valeurs (possiblement multiples) parmi les entités ayant pour type *séquence* ou l'une de ses spécialisations.

Cette approche peut se rapprocher d'une approche antérieure de représentation, qui partait des mêmes préoccupations d'adéquation cognitive à des utilisateurs humains et au fonctionnement supposé de leur mémoire : les réseaux sémantiques [Qui68]. Ceux-ci se concentrent sur la représentation des relations entre éléments conceptuels de manière graphique. Les connaissances sont en effet représentées sous forme de réseaux dont les nœuds sont des *concepts* (associables à des entités individuelles, prédicats unaires ou objets) et les arcs des *relations* entre concepts (associables à des prédicats binaires). On peut ainsi représenter notre index par le réseau de

[3]En général, on interprète la signification de la co-occurrence de notions dans un index comme une conjonction des significations respectives de ces notions : un index `{témoignage, professeur}` renverra vaguement à un thème complexe abordant ces deux sujets. Mais il s'agit plus d'une interprétation fondée sur l'habitude des usages que sur un engagement formel.

la figure 2.3. Dans un tel réseau, on constate bien que l'on définit l'émission considérée comme étant une[4] `Emission`, qui contient un objet rattaché à la notion d'`Interview`, qui est une sorte de `Séquence`, et que l'individu `prof_1`, de type `Professeur`, participe à cette interview.

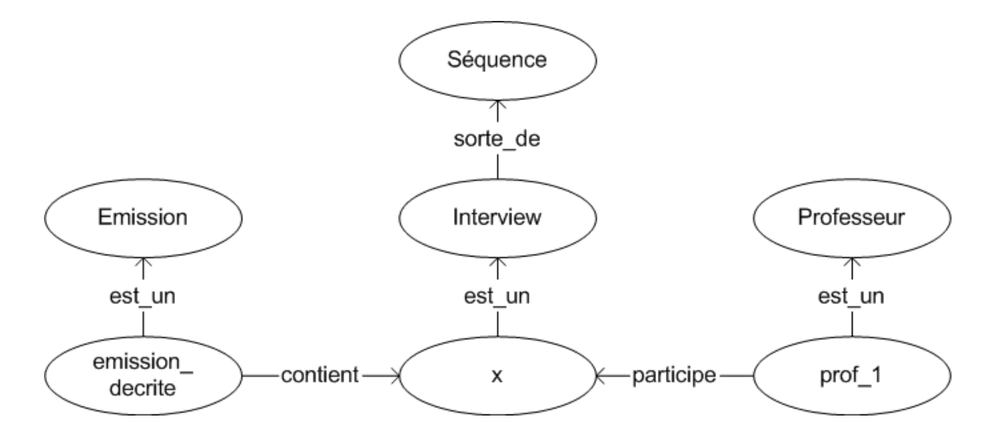


FIG. 2.3 Réseau sémantique représentant un extrait d'index

Sans contrôle au niveau de la création des réseaux, de telles méthodes de représentation présentent cependant des limitations immédiates. Il faut être rigoureux quant à la définition et au typage des liens employés, et faire la différence au sein du formalisme entre :

- les liens nécessaires à l'organisation rationnelle de la connaissance, pouvant faire l'objet de traitements spécifiques de la part d'un système de RC, comme la relation d'instanciation `est_un` ou la relation de spécialisation `sorte_de`, et
- les relations relevant uniquement du domaine d'application considéré, comme `contient` ou `participe`.

Il faut aussi éviter la confusion entre des informations relevant d'assertions accidentelles – les situations décrites – et celles relevant de connaissances essentielles – les descriptions valables pour toute situation envisageable dans l'application, comme la relation de spécialisation tenant entre la notion d'interview et celle de séquence dans un contexte audiovisuel. Et, finalement, on doit se poser la question des interprétations et des traitements qui sont possibles, en dehors de toute sémantique formelle. On peut évidemment exploiter des méthodes de recherche utilisant des algorithmes de traitement de graphes. C'est d'ailleurs ce qui a motivé en partie des approches comme les réseaux sémantiques, où l'on recherchait plus à exploiter la proximité entre les notions représentées dans le réseau qu'à raisonner sur une éventuelle interprétation « objective ». Mais ces algorithmes peuvent être d'une complexité trop importante, et leur application *ad hoc* ne garantit pas obligatoirement le passage, pour les informations manipulées, du statut de simple donnée à celui de connaissance. Et ceci vaut autant pour des systèmes qui auraient pour ambition de simuler des raisonnements que pour les utilisateurs eux-mêmes, dont l'interprétation des éléments rencontrés doit être guidée si l'on veut qu'ils exploitent correctement les informations fournies par un SBC.

[4] La relation d'instanciation – parfois désignée en RC par l'expression `est_un` – rattache un individu d'une description factuelle à une notion plus abstraite, *concept* (ou *classe* si on se place par exemple dans un paradigme de représentation objet), qui permet de le typer. Pour une discussion sur cette relation fondamentale, ainsi que d'autres, dans un contexte plus linguistique et cognitif, on peut se référer à [Des87].

Pour remédier à ces problèmes, il faut se tourner à nouveau vers des formalismes, mais des formalismes qui proposent des primitives adaptées à la modélisation de connaissances, ainsi que des algorithmes dédiés à l'exploitation spécifique des expressions composées à l'aide de ces primitives.

On trouve dans cette catégorie les *graphes conceptuels* proposés par John Sowa [Sow84], formalisme retenu dans le cadre du projet Opales. Ce langage de représentation propose une approche de la représentation logique qui est compatible avec l'utilisation d'interfaces graphiques de type réseaux sémantiques et exploite des algorithmes de raisonnement dont la finalité n'est pas la démonstration générale des formules du premier ordre (cf. section 2.3.2). Concrètement, un graphe conceptuel est un graphe biparti : les entités y sont de deux sortes distinctes, *concept* ou *relation*. Les sommets de nature conceptuelle sont étiquetés à la fois par un *type de concept* (correspondant à un prédicat unaire en logique) et par un *marqueur* individuel – possiblement *générique* – de sorte à dénoter un objet, connu ou inconnu, du domaine d'application. Les sommets de nature relationnelle – étiquetés par des relations – permettent de faire le lien, *via* les arcs du graphe[5] entre ces individus, comme les prédicats d'arité multiple rapprochent plusieurs objets en logique du premier ordre. Par exemple, un index de notre émission (cf. page 29) pourra être le graphe représenté en figure 2.4.

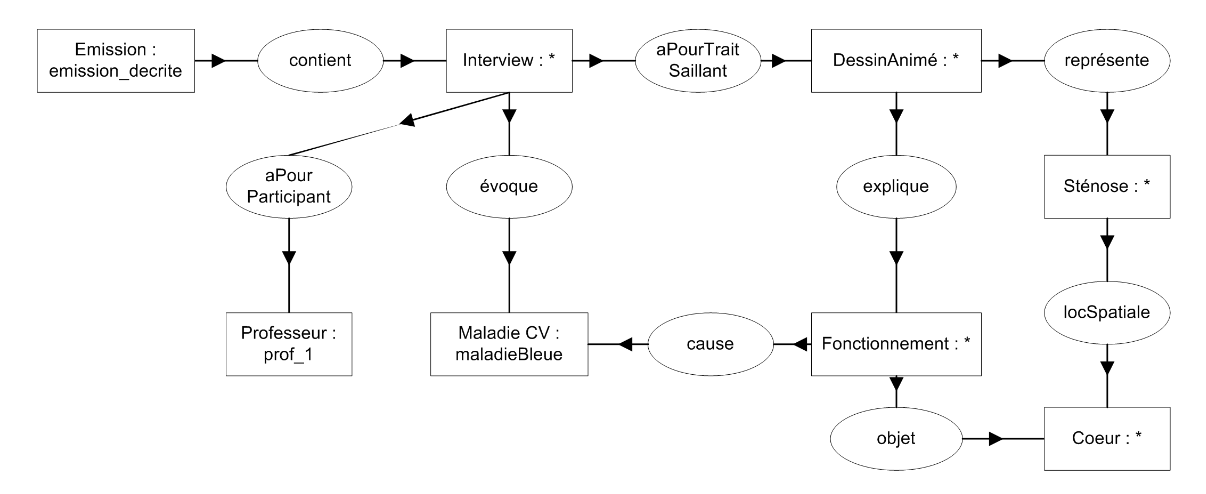


Fig. 2.4 Représentation d'un index sous forme de graphe conceptuel

*Le marqueur * dénote un individu indéterminé, rapprochable des variables introduites en logique du premier ordre par le quantificateur existentiel ∃.*

Un graphe conceptuel *stricto sensu* est tourné vers la description de cas particuliers observés dans le monde applicatif, et non vers la définition des notions générales qui sont employées pour décrire ces cas. Les différents types de concepts et les relations disponibles pour la création d'un graphe, ainsi que l'ensemble des marqueurs individuels apparaissent à part, dans ce que l'on appellera un *support*. C'est ce support qui renferme les liens structurant le domaine lui-même, comme l'information selon laquelle le type `Interview` spécialise le type `Séquence`, et qui sert de spécification pour les raisonnements possibles, comme nous le verrons dans les sections 2.2.2, 2.3.1 et 2.3.2.

[5]En principe, ces arcs sont numérotés pour ordonner les arguments d'une relation, mais dans le cas le plus fréquent, celui d'une arité relationnelle égale à 2, on préfère une notation employant des flèches.

Cette séparation entre connaissances du domaine d'application et connaissances liées aux cas particuliers représentés est partagée par une approche plus tournée vers la logique, celle des *logiques de descriptions* (LD)[6] [BCM+03]. Celles-ci dérivent directement de travaux comme ceux de Brachman [Bra79, BMP+91] qui ont distingué par exemple les liens *épistémiques*, appartenant au niveau de l'organisation formelle de la connaissance, des liens *conceptuels* et *linguistiques*, plus spécifiques de domaines ou de tâches particulières. Les LD sont inspirées par le paradigmes des *frames* et des réseaux sémantiques, mais elles reprennent la sémantique ensembliste formelle définie par la logique du premier ordre. Elles introduisent néanmoins une certaine quantité de *constructeurs* spécifiques – le nombre et la nature des constructeurs changent en fonction de la logique de description considérée – permettant :

- de contrôler l'expressivité autorisée pour les formules, et de se concentrer sur les primitives utiles pour une déclaration satisfaisante des informations au niveau de la connaissance ;
- de structurer facilement les connaissances ainsi déclarées, et de les répartir entre ce qui relève d'une conceptualisation générale à un domaine et d'assertions particulières, mais aussi de distinguer les différentes sortes de notions rencontrées à chacun de ces niveaux ;
- de s'appuyer sur cette structuration des connaissances et sur les restrictions opérées par rapport à la logique du premier ordre pour proposer des processus de démonstration plus efficients, voire complets.

Pour aider à la conception et à l'utilisation des SBC, les LD font clairement le distinguo entre des connaissances *terminologiques* – on parle parfois de *T-box* – définissant *concepts*[7] et *rôles* (nous verrons plus en détail dans la section 2.3.1 quelles sont les différentes possibilités en la matière) et des connaissances *assertionnelles* – *A-box* – décrivant les individus effectivement rencontrés dans les situations représentées, et comment ceux-ci sont reliés entre eux. Ces assertions sont semblables aux formules de la logique du premier ordre faisant intervenir des constantes nommées, à ceci près que les prédicats ne peuvent être qu'unaires ou binaires, puisque telle est la nature des concepts et rôles qui jouent un rôle prédicatif en logique de description. Une *A-box*, quelle que soit la syntaxe retenue, contiendra ainsi des expressions de la forme `C(a)` ou `r(a,b)`, où `C` et `r` désignent respectivement un concept et un rôle, et `a` et `b` font référence à des individus. Dans une logique de description autorisant un langage d'assertion minimal, on pourra par exemple représenter notre index par les assertions de la figure 2.5 : un objet particulier de type `Emission` contient un objet particulier de type `Interview` auquel participe un individu particulier de type `Professeur`.

```
Emission(emission_decrite)
Interview(interview_1)
Professeur(prof_1)
contient(emission_decrite, interview_1)
participe(prof_1, interview_1)
```

Fig. 2.5 Extrait d'A-box de logique de description simple, représentant un extrait d'index

[6] Ces formalismes, s'ils n'ont pas été retenus dans nos expérimentations initiales guidées par Opales, sont cependant à considérer, autant du fait de leurs qualités intrinsèques que de l'engouement qu'ils suscitent, comme on va le voir, dans la communauté de recherche du *web sémantique*. De fait, nous avons veillé à garder nos propositions compatibles autant que possible avec leurs spécifications (section 4.3.3) et les avons employées dans des expérimentations plus tardives (section 5.2).

[7] Ici, un concept renvoie à un prédicat unaire, non instancié – en logique, on emploie la formule introduite par Frege de prédicat *non saturé*.

Ici, on est obligé d'introduire des objets nommés, comme `interview_1` pour faire référence à chaque individu auquel la A-box fait référence. On se heurte ici à l'une des limitations expressives possiblement introduites par les LD : on ne considère que des constantes correspondant à des objets identifiés, on ne peut introduire de variable. Néanmoins, avec une logique plus expressive permettant l'emploi dans la A-box de descriptions complexes pour les instances de concepts, on pourra exprimer notre index avec plus de subtilité (figure 2.6), et garder l'« information » selon laquelle notre interview n'est pas désignée explicitement.

```
(Emission ⊓ (∃contient.(Interview ⊓ (∃participe⁻.(ONE_OF prof_1)))))
    (emission_decrite)
Professeur(prof_1)
```

FIG. 2.6 Extrait d'A-box de logique de description évoluée, représentant un extrait d'index

Les notations des constructeurs empruntent la syntaxe abstraite employée dans [BCM+03]. Ici, on a voulu exprimer le fait que l'émission décrite est une émission qui contient (au moins) une interview à laquelle participe au moins une instance du concept anonyme défini – par extension – comme contenant l'unique individu `prof1`*, qui est de type professeur.*

On a donc vu que les principaux langages de représentation de connaissances autorisent la création de ressources structurées. Même dans le cadre des langages les plus contraints, il reste toujours possible d'exprimer – et comme on verra par la suite, d'utiliser pour des raisonnements – des liens spécialisés entre les entités de la base de connaissances. Seuls varient le statut accordé à ces liens, entre entités de premier ordre et attributs d'une entité conceptuelle conçue comme l'élément central de la représentation, et l'arité permise, uniquement binaire ou bien quelconque. Dans un système documentaire reposant sur l'utilisation d'un langage de représentation de connaissances structurées, les index pourront donc bénéficier d'une capacité expressive supérieure à celle des systèmes documentaires traditionnels.

Au sein de la fonctionnalité **F1** introduite dans le chapitre 1 pour traiter ce problème d'expressivité, nous avions distingué les besoins concernant la forme des descriptions – la possibilité de créer et d'exploiter des expressions dont la forme est élaborée – de ceux relevant de leur substance – la possibilité de faire reposer ces expressions sur un vocabulaire riche et adapté au domaine d'application. S'agissant de l'expressivité formelle des index, les connaissances créées pour représenter les index s'inscrivent bien dans le paradigme relationnel que nous recherchions. Quel que soit le choix de représentation retenu pour ces liens – relations, propriétés, attributs ou rôles – la démarche de la représentation des connaissances permet une articulation riche des descripteurs introduits dans les index.

La situation, on va le voir, reste cependant plus délicate en ce qui concerne la substance des descriptions. Les langages de RC offrent en effet des constructeurs permettant de représenter des situations complexes, représentations qui peuvent, le cas échéant, servir d'index. Mais ces primitives *logiques* sont plus orientées vers la satisfaction de critères indépendants des domaines d'application : les expressions formées seront valides d'un point de vue syntaxique, et s'inscriront dans un cadre épistémologique bien défini. Il reste à rattacher à ce langage un *vocabulaire* de primitives *non logiques*, adaptées à la représentation de connaissances dans un domaine spécifique.

En principe, quel que soit le formalisme retenu, il est permis d'introduire autant d'étiquettes de prédicats, concepts, relations, classes, rôles ou autres que l'on a besoin. On peut donc agir sur les primitives *non logiques*, la substance des expressions construites avec le langage de RC,

et faire en sorte que celles-ci créent bien un cadre de représentation adapté au domaine visé. Un concept *Séquence* pourra tout naturellement être introduit pour catégoriser les différents individus dénotant les séquences du monde décrit. Cependant, pour que de telles primitives constituent un véritable *système* de représentation intelligible et exploitable, il faut soigner tout particulièrement leur introduction et leur emploi dans le langage, que ce soit pour la conception ou l'utilisation du SBC. L'objectif est bien de pourvoir un ensemble de descripteurs d'une interprétation qui soit utilisable dans un langage de construction de descriptions complexes, et que la signification qui est associée de ce fait aux expressions produites s'inscrive correctement dans le fonctionnement d'un système d'accès aux connaissances – en particulier en spécifiant les calculs qu'il peut effectuer. On rejoint là les considérations liées à la notion de *vocabulaire* de description dans les langages documentaires, qui imposent une conception et une articulation précises, entre langage de description, descripteurs et traitements applicables. Pour résoudre de tels problèmes, l'approche de la RC, suivie par l'initiative dite du *web sémantique*, propose le recours à des spécifications conceptuelles pouvant servir de vocabulaire de représentation : les *ontologies*.

2.2.2 Vocabulaire de représentation et ontologies

Dans le cadre des systèmes documentaires, on a vu qu'il faut préciser ce qui constitue la *substance* des descriptions. On doit fixer un vocabulaire dont l'emploi lors de la construction des expressions garantit à celles-ci une intelligibilité et des traitements adaptés à l'application documentaire visée.

Il en va de même en RC. Pour obtenir le statut de connaissance, que de nombreuses définitions introduisent comme celui d'une information à laquelle s'ajoute un savoir relatif à son utilisation dans un contexte précis, les données doivent utiliser un *vocabulaire* – un schéma, si l'on veut faire une analogie avec les systèmes de gestion de bases de données relationnelles – approprié. En effet, les SBC fonctionnent sur le principe que les données qu'il manipulent et les traitements qu'ils effectuent sont fidèles à une représentation théorique du domaine. Cette représentation définit un contexte qui permet d'abstraire de l'observation des situations rencontrées dans ce domaine des cas modélisés, exploitables par un système informatique. La cerner le plus correctement possible, puis la transposer en un ensemble de spécifications utilisables pour conduire des calculs sont donc deux activités importantes. Elles constituent le cœur de ce que l'on appelle l'*ingénierie des connaissances*, discipline qui mobilise tour à tour, comme on le verra dans les chapitres à venir, la linguistique, la logique, l'informatique ou même la philosophie[8] afin de construire des SBC efficaces dans leur contexte applicatif. Elle se concentre en particulier sur la conception des ontologies que nous avons déjà plusieurs fois évoquées.

Les ontologies : inspiration et objectifs

Historiquement, l'*Ontologie* est une discipline de la philosophie dont l'objet est l'étude systématique de la nature et de l'organisation de l'être. Le terme a été réutilisé dans une acception distincte, moins ambitieuse, en intelligence artificielle, où l'« être » est plus facile à appréhender, puisqu'il s'agit de ce que l'on choisit de représenter dans un système. Les *ontologies* y désignent des artefacts – au sens de construction artificielle – élaborés lors de la modélisation[9] conceptuelle,

[8] Pour plus de détails, le lecteur pourra se tourner vers les contributions récentes de [Cha03] et [Bac04a].

[9] Bruno Bachimont dans [Bac96] définit un modèle de connaissance comme un

> système de connaissances permettant de raisonner à propos d'une réalité et d'en anticiper l'évolution sous un certain point de vue.

et dont l'objectif est de jouer le rôle de référentiels conceptuels pour les SBC.

Cela fait un certain temps que l'intelligence artificielle a dégagé le besoin de créer des modèles résultant d'un effort d'abstraction du monde de l'application. On veut rendre compte de compréhensions partagées, que ce soit pour améliorer l'interopérabilité des SBC, pour augmenter le potentiel de réutilisation des éléments spécificiant iceux, ou plus simplement pour faciliter de manière générale l'accès et la création des connaissances qui y sont représentées. Mais si les chercheurs en ingénierie des connaissances semblaient d'accord pour adopter en la matière une définition et des objectifs compatibles avec l'activité de construction de SBC, encore fallait-il trouver un consensus.

Après quelques hésitations, une formule définitoire a fini par être largement utilisée ; il s'agit de celle énoncée par Thomas Gruber dans [Gru93] et [Gru95] :

une ontologie est une spécification explicite d'une conceptualisation

Cet énoncé reste cependant problématique : il faut en particulier savoir ce que l'on entend exactement par « conceptualisation », et, comme on le verra en section 2.3.1, cela ne se fait pas sans difficulté. Si la définition de Gruber a été facilement reprise, c'est avant tout parce qu'elle reste vague.

Pour préciser le sens de la notion d'ontologie de manière pratique, Bruno Bachimont propose de repartir des besoins réels des systèmes de RC. Comme pour les systèmes documentaires classiques, on a effectivement besoin d'un langage formel de représentation dédié à l'application. Dans le cas d'un SBC, il s'agit de présenter les concepts et relations permettant de décrire les cas, de créer une base de connaissances qui contient des assertions exploitables. Ces langages, on l'a déjà vu, utilisent deux sortes de primitives :

- les primitives logiques, qui ont une signification fixée par une sémantique formelle (ce sont les symboles des connecteurs logiques du calcul des prédicats, par exemple) ;
- les primitives non logiques : ce sont celles, propres au domaine, qui sont issues du processus de modélisation conceptuelle (les symboles de prédicats, par exemple).

Le problème est que les primitives non logiques n'ont pas de signification définie *a priori* dans un langage artificiel de RC. Les concepteurs de SBC utilisent souvent des symboles qui sont des termes de la langue naturelle appartenant au champ lexical du domaine considéré, en espérant qu'une interprétation correcte les associera directement à ce qu'ils ont imaginé. On spécifiera ainsi qu'on peut utiliser les termes d'« interview » et de « séquence » comme symboles pour les primitives de représentation. Mais il faut rappeler que le sens des termes de la langue est fortement variable, et que par conséquent la signification de telles primitives reste floue. Une autre spécification aurait pu retenir la dénomination d'« entretien », proche d'« interview ». Il faut donc un moyen de guider l'interprétation des symboles utilisés, de donner un sens aux primitives non logiques du langage ; c'est le besoin auquel sont censées répondre les ontologies. [Bac00] propose la formulation suivante :

> *Définir une ontologie pour la représentation des connaissances, c'est définir, pour un domaine et un problème donnés, la signature fonctionnelle et relationnelle d'un langage formel de représentation et la sémantique associée.*

Introduction des vocabulaires spécialisés dans des langages de RC

Par exemple, si l'on veut représenter un index sous la forme d'un graphe conceptuel, comme dans OPALES, il faut introduire des *types de concepts*[10] et des *relations* qui constitueront la partie du vocabulaire de représentation relevant du domaine d'application, le *support* des futurs graphes conceptuels. Pour notre exemple, on pourra introduire évidemment des concepts comme `Sequence`, `Personne`, mais aussi des types plus généraux du domaine audiovisuel, comme `ObjetAV` (regroupant les séquences et les programmes) et des relations comme `participe`.

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<!DOCTYPE CGXML PUBLIC "-//COGITANT//CGXML Format Specification 1.0//EN"
                       "http://cogitant.sourceforge.net/cgxml.dtd">
<CGXML xmlns:ina="http://www.ina.fr">
<support name="Exemple AV">
  <conceptTypes>
    [...]
    <ctype id="id-101240791089724631510348203210" label="Personne"/>
    <ctype id="id-101240791089724631510498354554" label="Sequence"/>
    <ctype id="id-101240791089724631510498398264" label="ObjetAV"/>
    [...]
  </conceptTypes>
  <relationTypes>
    [...]
    <rtype id="id-101240791089645150322641047820" label="participe"/>
    [...]
  <relationTypes>
</support>
</CGXML>
```

CODE 2.1 – Extrait d'un support de graphes conceptuels en CGXML.

Ce support joue un rôle fondamental par rapport au langage de représentation que constitue le formalisme des GC. Tout d'abord, concepts et relations y sont en effet présentés sous la forme de hiérarchies de spécialisation qui organisent les notions, des plus générales aux plus particulières – nous verrons plus en détails la signification de cette hiérarchisation, à la fois semblable et différente de celle présente dans les thesaurus. Ensuite, les mécanismes de construction de graphes valides ne permettent pas d'introduire de nouveaux types de concepts ou de relations lors de ce processus. On peut donc bien affirmer que le formalisme de GC se donne les moyens de définir de façon explicite les primitives non logiques propres à une application. Et qu'en ce qui concerne l'encodage d'index sous forme de graphes conceptuels, comme dans OPALES, on commence à entrevoir une sorte d'engagement ontologique : on se donne les moyens de créer et d'utiliser un vocabulaire de description propre au domaine concerné. Pour chaque application, on pourra introduire un nouveau support, susceptible d'apporter les notions nécessaires à la représentation de connaissances structurées dans ce cadre. Le code 2.1 montre comment on peut utiliser un langage concret – ou plutôt un méta-langage – de représentation de connaissance pour spécifier

[10]Par la suite, et sauf cas contraire mentionné explicitement, nous emploierons de manière unifiée le terme « concept » pour désigner ces « types de concepts », et comme en logique de description ou en logique du premier ordre, le terme « individu » pour désigner les anciens « concepts », c'est-à-dire les instances apparaissant dans les graphes conceptuels.

un tel vocabulaire. En l'occurrence, cet exemple utilise CGXML, introduit par David Genest pour le système COGITANT utilisé dans OPALES [GS98]. Les types de concepts et de relations y sont respectivement introduits par les balises XML `ctype` et `rtype`.

Des mécanismes semblables d'introduction de primitives de représentation non logiques sont utilisés dans les spécifications terminologiques des logiques de description que nous avons abordées en section 2.2.1. Le rôle de la *T-box* est précisément de présenter les concepts et les rôles – la *terminologie*, au sens des LD – qui seront utilisés pour construire les assertions de la *A-box*, chargées de représenter les cas rencontrés dans l'application – en ce qui nous concerne, les index. Pour notre exemple, une *T-box* appropriée à la création de la A-box de la figure 2.5 introduira des concepts comme `Sequence`, `Personne` et des rôles comme `participe` et `contient`. Les procédés de construction des bases de connaissances terminologiques assurent après que les primitives employées dans les assertions ont bien été introduites auparavant dans la terminologie. Dans le cas des LD, c'est la *T-box* qui joue le rôle comparable à celui d'une ontologie du domaine applicatif.

On a donc en RC des mécanismes qui permettent de construire des bases de connaissances s'appuyant explicitement sur des vocabulaires dédiés à des applications particulières. Si on adapte au cas de l'indexation ce genre de démarche, on est donc à même de traiter le problème de la substance des expressions représentant le contenu des documents. Dans la mesure où l'on choisit un formalisme de représentation de connaissances laissant les concepteurs du SBC libres de spécifier le vocabulaire de description, on peut donc espérer que celui-ci pourra être pertinent pour le cadre applicatif retenu.

Cependant, on a vu dans le chapitre 1 que la création d'un langage de description documentaire exigeait plus que la simple donnée d'un vocabulaire, fût-il aussi spécialisé que l'application le demande. De fait une approche aussi simple serait en retrait par rapport aux définitions évoquées pour les ontologies. Ces définitions, parlant de « conceptualisation » de « représentation » ou encore de « sémantique » se placent clairement du côté d'un engagement quant au sens des primitives créées. Ce problème rejoint tout naturellement celui de la continuité sémantique. Il faut que ce langage fournisse des solutions à même de résoudre ce problème dans le cadre d'un SBC. Comment garantir une interprétation constante et correcte des index créés avec les primitives introduites ? Peut-on contrôler l'utilisation du langage par les utilisateurs ? Peut-on garantir une exploitation des connaissances par le système qui soit cohérente avec ce que l'on attend dans le cadre interprétatif défini par l'application ?

En ce qui concerne le contrôle des traitements effectués par le système, par exemple, on se doute qu'on peut créer des spécifications formelles riches et adaptées à des traitements qui répondent à des besoins précis. Encore faut-il que, dans une approche à base de connaissances, ces traitements puissent être spécifiés de manière déclarative, au niveau des éléments conceptuels rencontrés dans le domaine plutôt qu'à celui des structures informatiques « concrètes » manipulées par le système.

2.3 Ontologies, contrôle et traitements pour des systèmes d'indexation et de recherche

De fait, l'emploi de SBC et d'ontologies apportent des réponses à toutes ces questions. Plus que constituer un simple vocabulaire, une ontologie a pour ambition de servir de spécification pour un *système* de représentation et de manipulation de connaissances, en accord avec une vue abstraite du domaine applicatif, accord qui lui confèrerait une légitimité au regard des usages

du domaine. Nous allons voir dans quelle mesure une telle approche peut être réalisée, et les possibilités que cela induit en matière de contrôle et d'assistance à l'utilisation d'un SBC, en vue d'une indexation.

2.3.1 Spécifications ontologiques et respect de la continuité sémantique

Nous avons vu dans le chapitre 1 qu'une fonctionnalité indispensable d'un système d'indexation est le besoin de contrôle, dans le sens de la continuité sémantique, qu'il s'agisse de la création des index ou bien de leur interprétation. Non seulement on a besoin d'un vocabulaire adapté à l'application, mais ce vocabulaire doit pouvoir fonctionner comme un *référentiel de sens* qui prescrit une interprétation et un emploi précis pour les descripteurs proposés. Ainsi, il doit pouvoir contribuer à la continuité sémantique qui doit être vérifiée dans la chaîne documentaire. En effet, un système informatique d'indexation et de recherche peut très bien être un composant indépendant du reste des processus liés à l'application qui lui sert de cadre. Par exemple, dans le cas de la réutilisation d'éléments documentaires dans de nouvelles productions, ça n'est pas lui qui va procéder au montage, mais bien un opérateur humain, qui doit avoir la certitude que les résultats renvoyés par le système sont bien cohérents avec son interprétation.

Cette fonctionnalité, **F2**, s'assure donc qu'il existe une signification des index qui soit intelligible par les utilisateurs et exploitable par les machines, et que la création et l'exploitation de ces index se font conformément à ce qui est attendu au regard de cette signification. Là encore, le recours aux ontologies peut apporter une solution.

En ce qui concerne la substance des descriptions, tout d'abord, les ontologies, si elles proposent un vocabulaire de termes que le système d'indexation s'engagerait à utiliser, font plus que cela, puisque selon la définition communément acceptée elles doivent rendre compte d'une *conceptualisation* d'un domaine. Cette notion reste cependant très floue, et il faut voir ce en quoi consistent concrètement ces constructions qui permettent de « spécifier une conceptualisation » et de faire en sorte que le SBC fonctionne conformément à un « engagement ontologique ». Le vocabulaire de représentation dont nous avons besoin doit en effet s'appuyer sur un effort d'abstraction et d'implémentation, dont l'objectif est la légitimation du fonctionnement du SBC au regard du domaine d'application. Si on reprend l'articulation entre descriptions, langage et système de la fin du chapitre 1, il s'agit d'ancrer les descripteurs du langage d'indexation dans un contexte interprétatif qui fasse en sorte que les descriptions soient exploitées par le système en accord avec les besoins de l'application.

Conceptualisations et engagement ontologique

La confusion originelle quant à la définition des ontologies en IA a engendré une très grande variété des produits ainsi désignés. L'enjeu le plus clairement identifiable de la création de ces artefacts étant la création d'un vocabulaire permettant de décrire des connaissances, toute liste de symboles plus ou moins structurée est susceptible de se voir – parfois très abusivement – répertoriée comme une ontologie.

Un thesaurus, par exemple, est un vocabulaire de description qui a fait l'objet d'une définition et d'une structuration recherchant explicitement le consensus et la réduction des ambiguïtés dans l'application informationnelle visée. Il peut également être utilisé tel quel dans un système d'information capable d'exploiter la hiérarchie de notions qu'il contient. Certains seront donc tentés de l'avancer comme une ontologie.

Nous avons cependant montré que les thesaurus, s'ils peuvent satisfaire les besoins de certaines applications, présentent des insuffisances majeures en termes de précision et de cohérence, ce qui

empêche les notions qu'ils apportent de fonctionner comme des connaissances exploitables par un système d'information sans perte de précision. Pour nous, il faut que les référentiels de sens que vont constituer les ontologies soient plus complets, et que les significations des primitives soient plus explicites et mieux intégrées dans le SBC.

Pour cela, on doit d'abord savoir ce qu'il faut introduire dans les ontologies pour qu'elles fonctionnent comme des spécifications convenables. Peut-on clarifier ces notions d'abstraction, de conceptualisation, de spécification, en accord avec les besoins que nous avons dégagés ?

Comme le signale Nicola Guarino dans [GG95] ou [Gua98a], le premier des problèmes est qu'il existe au moins deux définitions d'une conceptualisation qui, si elles sont antagonistes, restent toutes deux utilisables dans l'énoncé de Gruber.

Le premier sens du terme renvoie pour un domaine donné à l'ensemble des « objets, concepts, et autres entités dont on affirme l'existence dans un un cadre donné, ainsi que des relations qui existent entre eux ». En d'autres termes, « une conceptualisation est une vue abstraite et simplifiée du monde que nous voulons représenter pour un but donné ». Ces citations extraites de [Gru95] montrent que l'on se préoccupe bien ici de modélisation, mais d'une modélisation du monde en tant que configuration particulière d'objets, et non de celle d'un domaine, supposée s'appliquer à toutes les configurations possibles. Le problème vient du fait que la définition énoncée par Gruber ne se préoccupe que d'extension : elle considère directement les objets du monde, auxquels elle attribue des propriétés. Par exemple, on saura que pour une configuration donnée l'extension d'un concept *Séquence* va contenir trois individus, a, b et c, et que a et b seront reliés par la relation *contient* qui est définie par la donnée d'un ensemble auquel appartient le couple *(a,b)*. L'une des conséquences de cette définition est que si le monde modélisé change, alors la conceptualisation change elle aussi : une conceptualisation, dans ce contexte, revient à un modèle de connaissance valable pour un *cas* particulier, et non pour un domaine applicatif dans toute sa généralité.

Guarino, jugeant cette approche non satisfaisante pour rendre compte de la signification réelle des concepts utilisés comme vocabulaire du langage de RC, propose une autre définition. Celle-ci s'intéresse davantage aux propriétés des concepts présentés et à la manière dont ils réfèrent aux individus du monde, et ce pour toute configuration envisageable. Il s'agit plus de dégager des propriétés invariantes de ces concepts : des règles qui permettent, pour une configuration du monde donnée, de désigner les individus que l'on peut relier aux concepts. Ainsi, si l'on suppose données les concepts *Interview* et *Expert* et de la relation *participe*, on pourra préciser qu'*InterviewExpert* est le concept qui regroupe les instances d'*Interview* qui ont pour participant au moins un individu désigné comme *Expert*. S'il ne s'agit pas vraiment de définition en *intension*, au moins se préoccupe-t-on de degrés de définition plus abstraits que celui de l'énumération d'objets.

Pour Guarino, une conceptualisation associe à un vocabulaire de description un ensemble de mondes possibles, de configurations. Et spécifier une conceptualisation, c'est donner un ensemble d'axiomes permettant de contraindre l'utilisation du vocabulaire : pour une conceptualisation donnée, on va exclure les descriptions incohérentes, et par conséquent les mondes impossibles, en tout cas du point de vue de l'application informatique considérée.

Apparaît alors une autre difficulté. En effet, si pour Guarino une ontologie doit rendre compte de la manière dont l'utilisation d'un vocabulaire peut être en accord avec une conceptualisation (on parle alors d'*engagement ontologique*), il indique aussi que cela n'est pas entièrement possible. En effet, il montre dans [Gua98a] :

1. qu'une conceptualisation ne peut être appréhendée que partiellement *via* l'interprétation

des contraintes (qui ne sont pas des définitions) qu'elle exerce sur les différentes structures valides du monde de l'application – structures auxquelles Guarino fait référence en tant que *modèles*[11] – représentables en utilisant un vocabulaire donné ;

2. qu'une ontologie, qui pour Guarino consiste en la donnée d'une théorie logique finie, à savoir un ensemble d'axiomes, ne peut que réaliser une approximation des lois régissant tous les modèles possibles pour une application, modèles auxquels on n'a d'ailleurs pas forcément accès.

Une ontologie, pour Guarino, ne peut donc que refléter une conceptualisation de manière indirecte et toujours partielle. Et cela est relativement normal, puisqu'il propose d'utiliser comme spécification des axiomes qui ont une forme logique et une sémantique ensembliste. Une telle spécification reste de nature extensionnelle, et ne permet d'approcher qu'imparfaitement la complexité des significations généralement exprimées dans un domaine, en particulier de celles exprimées au moyen de la langue. Et ce, même si on utilise des mécanismes logiques élaborés permettant d'approcher une interprétation désignée comme « intensionnelle », tels les mondes possibles[12] de Kripke.

Des niveaux de spécification variés et complémentaires

A défaut de donner une spécification complète du sens d'un vocabulaire, une ontologie peut essayer de donner les règles et contraintes les plus à même de guider les interprétations qui sont faites, par les utilisateurs humains et le système, des connaissances représentées avec celui-ci. Le niveau de spécification à atteindre est variable, en fonction du degré d'exigence de l'application, et en particulier des traitements que l'on veut effectuer au moyen des index.

Tout d'abord, est commun à toutes les applications à base de connaissances le besoin d'informations qui spécifient réellement, pour l'humain qui les manipule, le sens des primitives utilisées. Les ontologies, à ce niveau, ne doivent pas nécessairement être écrites à l'aide d'un langage formel : l'important est la compréhension de ce qui est exprimé. C'est pourquoi beaucoup de spécifications sont données de manière informelle, voire superficielle : dans ce cas, on cherche seulement à présenter les termes d'une conceptualisation du domaine d'application. Certaines vont plus loin, en organisant et en détaillant quelque peu l'information livrée pour préciser dans le détail la signification des éléments du vocabulaire.

Le recours à la formalisation des ontologies est néanmoins indispensable si on veut les utiliser dans des SBC qui opèrent des traitements définis au niveau de la connaissance. Il faut alors alors préciser une interprétation des différentes primitives de représentation en utilisant des langages implémentés de représentation des connaissances, qui sont adaptés à l'expression de significations formelles. Comme nous avons commencé à le voir, il en existe de nombreux ; chacun d'entre eux a des capacités expressives spécifiques, et correspond donc à un besoin précis en termes de manipulations informatiques.

Le vocable « ontologie » a donc servi à désigner des constructions de complexité très variable, allant, pour un domaine quelconque, de la simple donnée d'un vocabulaire de ce domaine à celle de théories axiomatiques voulant préciser toutes les lois relatives aux prédicats logiques utilisés pour la représentation des connaissances relatives à ce domaine. [UG96] propose une catégorisation simple des ontologies selon le degré de formalisation, de « très informelle » à « rigoureusement

[11] Il s'agit ici de modèles au sens de la logique formelle, c'est-à-dire d'*interprétations* logiques qui vérifient les contraintes induites par un ensemble de formules.

[12] On pourra ainsi donner une caractérisation du lien entre *Sequence* et *Interview* par la formule $\Box(\forall x \; Interview(x) \rightarrow Sequence(x))$ qui indique que dans tous les mondes possibles du domaine applicatif, une interview peut être considérée comme une séquence.

formalisée » en passant par « semi-formelles », suivant que l'on ait affaire à des listes de termes de la langue, des hiérarchies structurées ou bien des axiomatisations logiques extrêmement détaillées.

In fine, une ontologie fournit donc toujours un vocabulaire, et une interprétation de ce vocabulaire susceptible d'aider – ou de contraindre – son emploi dans un SBC. La définition de Gruber peut toujours s'appliquer, à condition de prendre en compte la dimension centrale pour un SBC de *langage* : produire *une* spécification d'une conceptualisation, c'est fournir un vocabulaire dédié à une application et essayer de rendre compte, de la manière la plus explicite et fidèle possible, de ce à quoi on se réfère dans le monde visé lorsqu'on emploie ce vocabulaire. Pour cela, il faut expliciter des contraintes, formelles ou non, qui guident l'interprétation des notions proposées.

Pour conférer aux descriptions qui seront produites avec les notions apportées par une ontologie le statut de connaissances, il est important de ne pas oublier la dimension consensuelle, nécessaire à une ontologie. Même dans le cadre d'une application restreinte, l'ontologie doit être rattachée à une signification partageable, non ambiguë et compréhensible par tous les utilisateurs de l'application [Bac96]. Les spécifications que constituent les ontologies doivent donc être duales :

– *interprétation formelle et opérationnelle* : les notions doivent être représentées de manière à avoir une signification utilisable par un système. Les systèmes informatiques étant d'essence formelle, cette signification doit s'inscrire dans un système formel, tel que la logique. C'est l'approche qui a guidé les premiers travaux sur les ontologies, on verra en section 2.3.1 des exemples de langages offrant cette possibilité.
– *interprétation intelligible* : les notions doivent être compréhensibles pour l'utilisateur final dans le cadre de l'application visée. La modélisation doit donc recourir à une représentation graphiquement compréhensible, ou offrir une explicitation langagière de sa signification, par le biais de définitions ou d'annotations textuelles des ressources introduites.

De fait, les deux volets – formel et informel – sont indissociables dans la conception d'une ontologie, en tout cas si l'on s'en tient à l'acception désormais la plus couramment répandue de cette notion. Il faut néanmoins noter que si le recours au niveau formel est impératif, en principe on ne doit pas spécifier de ce côté plus de choses qu'on ne le fait en direction des utilisateurs humains. On peut avoir envie d'exprimer des informations dont l'ordinateur n'aura pas forcément besoin, mais, dans une approche à base de connaissances, l'ordinateur n'a pas à effectuer des traitements sur la base de renseignements que l'on n'aura pas explicitement déclarés au niveau des connaissances.

On peut essayer de collectionner tous ces éléments définitoires, pour rappeler qu'une ontologie fournit un vocabulaire de représentation utilisable au sein d'un SBC, par le biais d'une spécification explicite, à la fois formelle et accessible, d'une conceptualisation à vocation consensuelle. Et nous renvoyons à nouveau à la figure 2.1 pour une vision d'ensemble de la place des ontologies dans les SBC.

Spécification et continuité sémantique

L'ontologie doit donc remplir deux rôles complémentaires : l'un est d'apporter des significations compréhensibles pour des humains, l'autre de fournir une spécification exploitable par les ordinateurs, la seconde devant être conforme à ce qui est affirmé dans les premières. C'est cette dualité, multipliant les champs d'application possibles des ontologies, qui a participé à leur succès.

Une ontologie consiste tout d'abord en la donnée d'un vocabulaire conceptuel. Elle devra par conséquent comprendre un ensemble de concepts étiquetés par des *symboles*, qui peuvent être des mots de la langue ou des symboles totalement arbitraires. Quelques ontologies ont pu se contenter de l'énoncé de tels symboles, en leur adjoignant quelquefois des définitions informelles. En étiquetant les primitives par des *termes*, en faisant référence à la langue naturelle et au vocabulaire qu'elle utilise dans le domaine d'application, on espère ainsi transposer une interprétation linguistique dans le contexte du SBC. Néanmoins, notre étude des systèmes documentaires a montré que ce genre de solution est imparfait, surtout dans un cadre d'indexation exigeant. Non seulement il faut se tourner vers des paradigmes de représentation riches, mais il faut prescrire pour le vocabulaire employé une signification précise, prescription qui est en mesure d'assurer la continuité sémantique.

Les thesaurus constituent, on l'a vu, une première approche pour régler ce problème de continuité. Ils organisent les notions employées dans le langage de description, ce qui leur confère un contexte d'interprétation intelligible. Cette organisation peut même être exploitée au sein de systèmes assistant les recherches, *via* des mécanismes de reformulation de requêtes. Mais on a vu que les spécification thésaurales sont par trop incomplètes, et les index construits, pas assez riches et précis.

Il faut donc aller plus loin, ce qu'autorisent justement les approches de représentation de connaissances. En effet, si on peut grâce au recours à des formalismes élaborés représenter des cas complexes, on peut également introduire les primitives des langages conçus avec ces formalismes en faisant en sorte qu'ils soient munis d'une signification précise, indépendamment même de leur emploi dans une base de faits.

organiser les connaissances Le premier moyen de définir des primitives de représentation de connaissances est de recréer un contexte d'interprétation pour chacune d'entre elle. A l'instar de ce qui se passe dans le cas de la langue naturelle, il est possible de faire ceci au moyen du langage de représentation lui-même, indépendamment de tout système de référence indicielle au monde « réel ». Il faut construire un véritable *système* symbolique, où les primitives de représentation sont organisées de manière à s'inter-définir.

Un procédé classique pour obtenir un tel réseau de significations est de construire un réseau mettant les primitives en relation les unes avec les autres par des liens organisationnels, que l'on sait interpréter hors de tout domaine. De fait, les langages de représentation de connaissances vont très souvent avec un certain nombre de constructeurs de nature épistémique – on peut alors parler de méta-langages, puisque ces constructeurs visent à la création des langages de représentation de connaissances factuelles. Ces constructeurs sont répartis en primitives permettant de déclarer :

- le type des primitives de représentation : *types de concepts/types de relations*, *concepts/rôles*, ou bien encore *classes/attributs*, pour mentionner les choix faits[13] dans les paradigmes les plus courants ;
- les liens épistémiques mettant en relation les primitives déclarées au moyen de ces types.

On retrouve là une démarche semblable à ce qui se fait dans les thesaurus, à deux différences fondamentales près :

- puisqu'on est dans des paradigmes complexes, on dispose de plus de constructeurs, à la fois pour les types épistémiques et pour les liens possibles entre ces types ;
- sous l'influence de l'intelligence artificielle et de sa tradition formaliste, les constructeurs sont munis d'une sémantique non ambiguë, la plupart du temps formalisée.

[13]Pour rester homogènes avec nos propres choix, nous privilégierons l'opposition concepts/relations.

Bénficiant de plus de possibilités expressives, et d'une précision bien plus grande quant à l'interprétation des liens épistémiques organisant le réseau de notions, on a donc les moyens d'éviter les faiblesses des spécifications thésaurales.

Ainsi, il est toujours souhaitable de créer, pour les primitives conceptuelles simples ou relationnelles, des hiérarchies de spécialisation : on définit une notion en référence à celle, qui plus générale qu'elle, lui fournit un contexte sémantique. Ces hiérarchies, apparues très tôt (dans la philosophie antique grecque, avec Aristote puis Porphyre), constituent en effet un moyen naturel d'organiser le vocabulaire d'un domaine et de faciliter sa compréhension. On autorise une comparaison immédiate – facilitée le cas échéant par une représentation graphique arborescente – entre une notion et ses voisines, ce qui facilite l'émergence d'environnements d'interprétation locaux. Par généralisation des comparaisons en suivant les liens de spécialisation/généralisation, on obtient ainsi, à l'échelle du domaine, un ensemble de significations inter-dépendantes [Bac96]. Les arborescences des figures 2.7 et 2.8, issues d'une ontologie pour l'audiovisuel conçue par Raphaël Troncy et nous-mêmes[14] [IT04], illustrent un tel agencement.

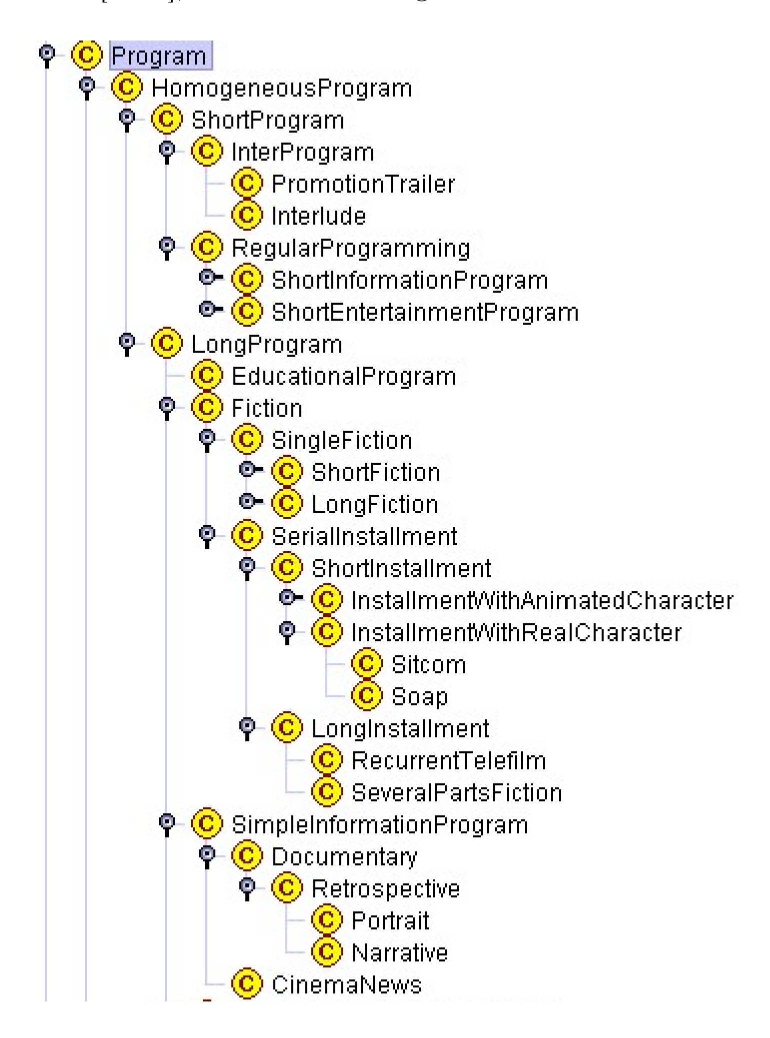


Fig. 2.7 Hiérarchie de concepts issue d'une ontologie pour l'audiovisuel, extrait de [IT04].

Cependant, dans le cadre de langages de RC formels, l'interprétation de ces liens de spécia-

[14] Plus de détails seront donnés en section 5.2.3.

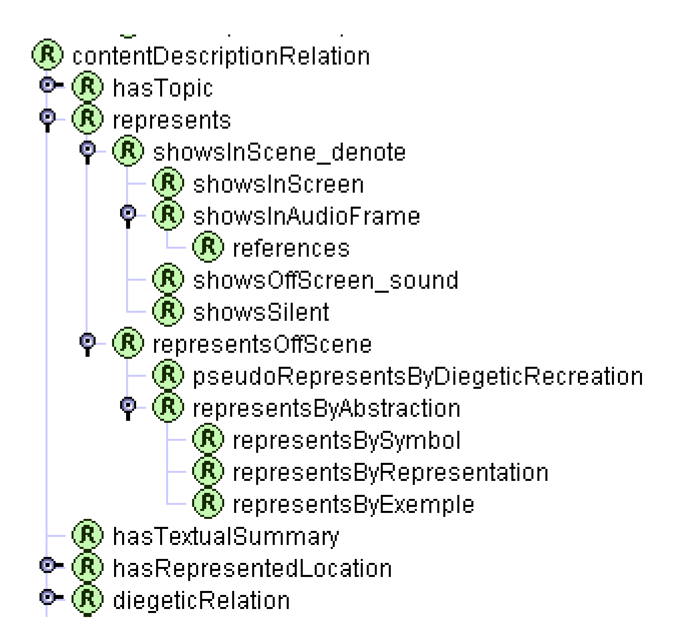


Fig. 2.8 Hiérarchie de relations conceptuelles issues d'une ontologie pour l'audiovisuel, extrait de [IT04].

lisation est très précise. Les primitives de représentation sont en effet généralement interprétées de manière ensembliste, les concepts correspondant à des ensembles d'individus du domaine applicatif. Dans un tel cadre, la relation de subsomption entre concepts est interprétée de manière stricte comme correspondant à la relation d'inclusion ensembliste. Si le concept `Interview` est introduit comme spécialisant le concept `Séquence`, cela implique que tous les éléments dénotés par le premier sont dénotés par le second.

On a donc bien une hiérarchie qui rend compte d'une seule sorte de décomposition des types d'entités du domaine. Il en résulte une plus grande homogénéité dans l'enchaînement des concepts généralisants et spécialisants : on ne peut plus mélanger, par exemple, la spécialisation taxonomique (de l'espèce à la sous-espèce) et la spécialisation méréonomique (du tout à la partie), comme c'était souvent le cas dans les thesaurus. En respectant de telles contraintes, on améliore ainsi le processus de modélisation conceptuelle, ainsi que son résultat. On prescrit des significations qui, plus précises, à la fois sur le plan de l'intelligibilité et sur celui de la formalisation, seront plus facilement ré-interprétables de manière convenable, par un utilisateur humain ou par un système de calcul formel. Ceci ne peut que faciliter le respect de la continuité sémantique.

RC formelle et graphes conceptuels Les GC que nous avons évoqués en section 2.2.1, ont par exemple été introduits comme des constructions associables à une signification formelle. Cette signification, exprimable en logique du premier ordre est un premier pas vers une opérationnalisation par des algorithmes de raisonnement des connaissances construites à l'aide des graphes.

Dans cette optique, les primitives – concepts et relations – introduites dans le support sont organisées par la relation de subsomption, et sont munies d'une interprétation ensembliste contrainte par cette relation. Les concepts dénotent des ensembles d'objets individuels, et les relations des produits cartésiens d'ensembles. Et d'un point de vue formel, la relation de spécialisation entre concepts est interprétée comme affirmant l'inclusion des ensembles dénotés par

lesdits concepts.

Il est à noter que l'interprétation ensembliste des concepts des GC n'interdit pas qu'un concept puisse spécialiser plusieurs concepts distincts dans le support : on peut ainsi imaginer que le cas échéant un concept `Interview` puisse spécialiser à la fois les concepts `Temoignage` et `Dialogue`. On parle alors d'héritage ou de hiérarchie *multiple*.

On a déjà vu que le support est aussi chargé de l'introduction des relations conceptuelles. Celles-ci sont également définies par une hiérarchie, mais en ce qui les concerne cette hiérarchie est complétée par la donnée d'un ensemble de « graphes canoniques ». Ces graphes mobilisent les types de concepts et de relations introduits dans le support pour créer des graphes élémentaires (composés d'une seule relation, comme en figure 2.9) qui permettent de construire les seuls graphes valides pour le SBC. En effet, dans le formalisme des GC, on construit les graphes d'une base de faits en effectuant des opérations à partir de graphes existants :

- *restriction* – spécialisation ou instanciation d'un concept, ou d'une relation, en suivant les informations hiérarchiques données dans le support ;
- *jointure* – réunion de deux graphes avec « fusion » de leurs concepts et relations communs ;
- *simplification* – suppression des relations redondantes entre concepts – et
- *copie*.

Dans un tel cadre, il faut des graphes « préexistants », servant de points de départ à la conception de nouveaux graphes : c'est justement le rôle de la *base canonique*, qui constitue le troisième élément du support. Dans le format CGXML, ce sont les attributs `idSignature` de la balise `rtype` (cf code 2.2) qui introduisent le graphe canonique associé à la relation concernée.



FIG. 2.9 Le graphe canonique associé à la relation `aPourParticipant`

Ce graphe présente de fait les concepts les plus généraux que la relation pourra mettre en rapport. On parle de *signature*, et par analogie avec les relations et fonctions mathématiques, de *domaines*, ou de *domaine* et de *co-domaine* dans le cas des relations binaires. Ces informations seront interprétées par les moteurs d'inférence GC, en conjonction avec les hiérarchies de spécialisation, pour valider les graphes produits. Ainsi, un outil de création de descriptions utilisant un service de validation conforme à la sémantique des GC refusera, en s'appuyant sur la définition de la relation `aPourParticipant`, toute assertion de cette propriété entre des instances de types de concepts n'étant pas ou ne spécialisant pas les types `Sequence` et `Personne`.

La création d'un support hiérarchique crée donc des contraintes interprétables à la fois par le concepteur ou l'utilisateur du SBC, contraintes qui ont également une signification pour le système d'inférence accédant à une version encodée de ce support. Toute spécification conceptuelle du vocabulaire devra donc tenir compte de cette interprétation, respectée par les mécanismes de raisonnement exploitant les graphes conceptuels en accord avec le support. Les moteurs d'inférence de graphes conceptuels reposent en effet sur la projection, opération qui vise à faire correspondre un graphe donné à un graphe qui le « spécialise » (en fonction des opérations de construction de graphes). Par exemple, on peut indiquer qu'`Interview` spécialise `Sequence`, comme le montre le code CGXML 2.2 qui utilise les balises `order`. Par la suite, un sommet conceptuel de type interview et de marqueur donné sera considéré comme une spécialisation d'un sommet de marqueur

```
<conceptTypes>
    [...]
    <ctype id="id-101240791089724631510498354554" label="Sequence"/>
    <ctype id="id-101240791089724631510498312578" label="Interview"/>
    <ctype id="id-101240791089724631510348203210" label="Personne"/>
    <ctype id="id-101240791089724631510498354554" label="Sequence"/>
    <ctype id="id-101240791089724631510498398264" label="ObjetAV"/>
    [...]
    <order id1="id-101240791089724631510498312578"
           id2="id-101240791089724631510498354554"/>
    <order id1="id-101240791089724631510498354554"
           id2="id-101240791089724631510498398264"/>
  </conceptTypes>
  <relationTypes>
    [...]
    <rtype id="id-101240791089645147127985121078"
           label="relationDescriptionContenu"/>
    <rtype id="id-101240791089645150322641047820" label="represente"/>
    <rtype id="id-101240791089645150322641047820" label="participe"
       idSignature="id-101240791089724631510348203210
                    id-101240791089724631510498398264"/>
    [...]
    <order id1="id-101240791089645150322641047820"
           id2="id-101240791089645147127985121078"/>
  <relationTypes>
```

Code 2.2 – Spécifications formelles dans un support GC en CGXML.

identique, mais de type séquence. Un moteur d'inférence calculera ainsi qu'un graphe indiquant qu'une personne donnée participe à une interview quelconque spécialise un graphe représentant un relation conceptuelle quelconque entre une personne quelconque et une séquence, et en tirera des conséquences significatives du point de vue des réponses retournées (cf section 2.3.2).

Cette approche revient donc bien à créer une ontologie formalisée d'un domaine. On distingue bien les types d'entités qui nous intéressent, et on les munit d'une signification à la fois précise et intelligible. Concepts et relations sont interprétables par la combinaison de leurs intitulés langagiers et de la relation de spécialisation. Dans le cas des relations, la donnée des concepts qui jouent pour elles le rôle de domaines vient compléter ces renseignements.

Toutes ces informations ont leur intérêt autant pour les utilisateurs humains que pour le système. Les spécifications sont en effet interprétées par le SBC, puisque des mécanismes de contrôle – dérivés de la sémantique formelle – sont mis en place, qui respectent la double sémantique des primitives. Ainsi un système à base de GC offre donc des garanties intéressantes quant à la continuité sémantique. Il est à noter que les significations qui sont spécifiées dans les supports permettent un contrôle qui concerne autant :

- la *substance* des descriptions : on a vu précédemment que les graphes ne contiennent que des concepts et des relations issus du support ;
- la *forme* des descriptions : signatures et hiérarchies, en conjonction avec le respect de règles de construction précises, permettent de contrôler les combinaisons de concepts et de relations qui sont considérées comme légales dans le domaine.

Expressivité formelle et continuité sémantique, l'exemple des LD On peut définir de manière similaire les primitives de représentation utilisées dans les logiques de description introduites en section 2.2.1. Là aussi, les spécifications terminologiques des *T-box* introduisent les concepts et relations – rôles, pour reprendre le vocabulaire des LD – au sein de hiérarchies interprétables formellement, et leur assignent des contraintes additionnelles qui précisent leur signification tout en restreignant leur emploi dans des bases d'assertions.

```
Interview ⊑ Sequence
Professeur ⊑ Personne
aPourParticipant ⊑ range(Personne)
aPourParticipant ⊑ domain(Sequence ⊔ Programme)
aPourInvite ⊑ aPourParticipant
RoleExpert ⊑ Role
RoleProfesseur ⊑ RoleExpert
joueRole ⊑ domain(Personne)
joueRole ⊑ range(Role)
```

FIG. 2.10 Extrait de T-box

Ainsi, on peut spécifier un concept (respectivement rôle) en tant que spécialisation d'un autre concept (resp. rôle). On peut également préciser, pour un rôle, le domaine – *domain* – et le co-domaine – *range*. On trouvera par exemple dans une *T-box* les axiomes de la figure 2.10, qui indiquent que les concepts `Interview` et `Professeur` sont respectivement des spécialisations – on utilise le constructeur $\sqsubseteq$ qui dénote l'inclusion ensembliste – de `Sequence` et `Personne`, et que le rôle `aPourInvite` est subsumé par `aPourParticipant`. Le rôle `aPourParticipant`, lui, relie des séquences ou des programmes à des personnes. Des connaissances semblables seront incluses pour caractériser les différents rôles que peut jouer une personne. On notera d'ailleurs

que l'association de `joue` à ses domaines est ici indispensable pour lever l'ambiguité sémantique qui résulte du choix d'une étiquette langagière hautement polysémique...

Ces spécifications, comme dans le cas des GC et de la logique du premier ordre qui les soustend, sont interprétées de manière ensembliste : les concepts dénotent des ensembles auxquels appartiennent les éléments les instanciant. Mais les logiques de desciption ont vocation à exprimer des définitions employant d'autres constructeurs que les GC. On peut vouloir dire par exemple que plusieurs concepts réalisent une partition d'un autre – toute instance de ce dernier est obligatoirement instance d'un des premiers – ou encore attribuer des propriétés algébriques à une relation binaire – réflexive, transitive, etc. Il est ainsi possible de construire des expressions complexes, au moyen de primitives épistémiques prenant en charge des besoins attestés dans nombre des applications de représentation de connaissances[15]. Les constructeurs utilisés dépendent en fait de la logique retenue. Suivant les possibilités de calculs et les impératifs de modélisation, on aura recours à des logiques plus ou moins expressives, de complexité algorithmique plus ou moins importante [BCM+03].

Parmi les constructeurs les plus fréquemment employés, on peut mentionner ceux qui ont été retenus pour le langage de représentation d'ontologies OWL [HPv03, OWL04]. Et plus particulièrement à son sous-ensemble OWL-DL, qui s'appuie sur la logique $\mathcal{SHOIN}(D)$ et comprend – entre autres – les primitives épistémologiques du tableau 2.1.

Définition de classes	*Définition de propriétés*	*Définition d'individus*
Class	ObjectProperty	Individual
subClassOf	subPropertyOf	type
equivalentClass	equivalentProperty	sameIndividualAs
disjointWith	domain	differentFrom
intersectionOf	range	AllDifferent
unionOf	inverseOf	
complement Of	SymmetricProperty	
oneOf	FunctionalProperty	
someValuesFrom	InverseFunctionalProperty	
allValuesFrom	TransitiveProperty	
hasValue	DatatypeProperty	
minCardinality		
maxCardinality		
cardinality		

TAB. 2.1 Constructeurs du langage OWL-DL

Avec ces constructeurs, on peut par exemple caractériser des concepts comme celui d'une séquence de dialogue, en introduisant une condition nécessaire qui spécifie qu'on a affaire à une séquence ayant au moins deux participants :

$$\texttt{SequenceDialogue} \sqsubseteq \texttt{Sequence} \sqcap (\geq_2 \texttt{aPourParticipant Personne})$$

FIG. 2.11 Caractérisation du concept `SequenceDialogue` en LD

[15]On pourra par exemple consulter sur le site du W3C les cas d'utilisation guidant la conception du langage OWL, proposé dans le cadre de l'initiative du web sémantique (cf. p. 2.4). Et lire les nombreux messages liés à ce sujet sur les listes de discussions afférentes.

Cette logique permet l'expression de relations d'équivalence entre concepts complexes : il est possible d'affirmer que tel concept spécialise tel autre, ou bien spécialise une expression complexe, mais aussi d'énoncer qu'un concept est spécialisé par une expression de ce type. On peut ainsi construire de véritables *définitions* formelles, apportant des conditions nécessaires et suffisantes de reconnaissance d'un concept ou d'un rôle dans une BC. Par exemple, on peut introduire le concept d'interview d'expert comme correspondant exactement aux interviews dont au moins l'un des participants est reconnu comme un expert (fig. 2.12). Le système pourra reconnaître qu'un objet instancie ce concept dès que les conditions de la définition seront vérifiées. Mais il pourra aussi considérer que tout objet instanciant ce concept induit l'existence des objets et des liens spécifiés dans cette définition. Ainsi, si une séquence particulière est décrite comme une interview d'expert, on pourra la renvoyer parmi les résultats d'une requête demandant des « séquences auxquelles participent des experts », même si cette information n'a pas été spécifiée dans l'index. On peut aussi redéfinir plus précisément les spécialisations de `Personne`, en indiquant quelle est la nature des rôles que les individus qui les instancient jouent (figure 2.12).

$$\texttt{InterviewExpert} = \texttt{Interview} \sqcap (\exists \texttt{aPourParticipant}\ \texttt{Expert})$$
$$\texttt{Professeur} = \texttt{Personne} \sqcap (\exists \texttt{joueRole}\ \texttt{RoleProfesseur})$$
$$\texttt{Expert} = \texttt{Personne} \sqcap (\exists \texttt{joueRole}\ \texttt{RoleExpert})$$

FIG. 2.12 Définition des concepts `InterviewExpert`, `Professeur` et `Expert` en LD

Un interview d'expert est un interview auquel participe au moins une instance du concept expert, et un professeur (respectivement un expert) est une personne qui joue un rôle de professeur (respectivement un rôle d'expert).

La spécification par une condition nécessaire – qui correspond à l'interprétation logique de la relation de spécialisation – permet de spécifier des propriétés essentielles de manière indirecte : on sait ce qu'appartenir à l'extension d'un concept implique pour l'individu qui l'instancie. Mais avec une définition par condition nécessaire et suffisante, il est possible à présent de définir explicitement ce qui permet d'identifier la notion, son essence. Si on reste dans le cadre d'une interprétation formelle, avec ce que tout cela peut impliquer de distance par rapport aux usages d'une application, il y a là une avancée importante vers la saisie – et l'utilisation dans des systèmes de raisonnement – de l'intension des notions constituant le vocabulaire d'un langage de RC. Et, partant, des possibilités bien plus grandes d'établir un contrôle sémantique précis, applicable à la fois lors de la création, du traitement automatique ou de l'accès aux connaissances.

Le recours à des formalismes de définition riches permet donc d'étendre considérablement l'efficacité des systèmes en termes d'interprétation des descriptions à base de connaissances constituant les index. Les ontologies apportent en fait une réponse crédible aux besoins de contrôle induits par l'exigence de continuité sémantique. Tout d'abord, le processus de définition « naturelle » des concepts et relations se voit amélioré par rapport à ce que permettaient par exemple les thesauri : associer par exemple une signification plus précise à la relation hiérarchique permet de clarifier le positionnement d'une notion par rapport à son contexte d'interprétation et d'aller plus loin dans le processus de sa désambiguïsation. De fait, l'accent mis sur la recherche de consensus et le partage des notions lors de la création d'une ontologie correspond tout à fait aux impératifs guidant la conception des vocabulaires de description : l'engagement ontologique renvoie d'une certaine manière aux choix qui étaient faits lors de la définition de langages documentaires traditionnels. A ceci près qu'on recherche pour les notions d'une ontologie des

spécifications plus complètes que celle des descripteurs thésauriques, au besoin en utilisant des définitions élaborées reposant sur des interprétations formelles. L'engagement plus strict dont bénéficient les ontologies, que ce soit de la part de ceux qui les conçoivent ou de ceux qui les utilisent, garantit une meilleure compréhension du contenu des descriptions, et donc favorise la continuité sémantique.

Ensuite, le fait que les descripteurs se voient à présent placés dans un cadre formel permet alors à un SBC, pour qui de telles spécifications sont accessibles *via* leur encodage dans un métalangage de RC muni d'une sémantique formelle, de jouer un rôle actif en ce qui concerne le contrôle de la continuité sémantique. Comme on l'a déjà évoqué pour le cas particulier des GC, Les SBC peuvent en effet invoquer des mécanismes de vérification automatique pour contrôler la validité des index produits, au regard des contraintes formulées au moyens de langages qui ont une signification opérationnelle adaptée. Les logiques de description apportent par exemple des mécanismes de satisfiabilité, qui vérifient la cohérence de la base d'assertions par rapport à ce qui est introduit dans la base terminologique de la *T-Box*. Les spécifications qui sont entrées dans les ontologies formalisées permettent alors au système documentaire de contrôler lui-même la continuité sémantique des index, tant au niveau de la substance des descriptions qu'à celui de leur forme.

Le passage à une interprétation formelle permet donc d'obtenir une plus grande cohérence en ce qui concerne l'utilisation des descripteurs composant les index, cette formalisation a des conséquences plus générales en termes d'exploitation. Si on reprend par exemple la classification des contextes d'usage proposée par [Für04], une ontologie contenant des définitions et des axiomes formalisés peut être en effet être utilisée à fins de *validation* et d'*inférence*[16]. De fait, une fois munis de spécifications formelles, les systèmes peuvent opérer des calculs qui, au-delà de la vérification et du contrôle sémantiques, peuvent être utilisés pour améliorer l'exploitation des index dans le système documentaire. Il s'agit alors d'une application plus « active » des connaissances formelles, dans le but de produire de nouvelles assertions dans la base de faits constituée par les index. Nous allons voir à présent comment l'utilisation de telles capacités de raisonnement est tout à fait pertinente pour un SBC à vocation documentaire.

2.3.2 Inférence, continuité sémantique et pertinence d'un SBC

Un cadre formel pour une assistance plus précise et plus riche à la recherche d'information

On a vu dans le chapitre 1 qu'un système documentaire devait proposer des fonctionnalités de manipulation des index permettant l'assistance à la recherche des informations ou des segments documentaires pertinents – notre item **F3** de la page 35. En particulier, il serait intéressant que le système prenne en charge une partie des reformulations dont on a vu dans ce même chapitre 1 qu'elles étaient souvent nécessaires à la recherche des index. Les indexeurs ne peuvent en effet anticiper tous les besoins à venir, et créent des index qui parfois ne contiennent qu'implicitement les informations nécessaires au traitement correct des requêtes. Ceux qui effectuent les recherches sont alors obligés de reformuler les requêtes pour tenir compte de ces lacunes. On peut faire remarquer que ces traitements peuvent être considérés comme s'inscrivant dans le cadre de la

[16]Guizzardi, dans une classification comparable, utilise les termes de *consolidation* et de *dérivation* pour caractériser les règles de raisonnement introduites dans une ontologie, suivant que ces règles vérifient la consistance d'une base de connaissance ou y ajoutent de nouveaux éléments [GFP02].

recherche de la continuité sémantique, puisqu'il s'agit de ré-interpréter les index effectivement exprimés à la lumière des requêtes posées.

Les systèmes documentaires classiques, on l'a vu, peuvent assumer eux-mêmes une partie de ce travail de reformulation en s'appuyant sur les connaissances du domaine. Par exemple, les relations hiérarchiques présentes dans le thesaurus sont exploitées de manière intensive par les systèmes, qui considèrent que l'on peut généraliser les descripteurs contenus dans un index pour répondre à une requête. Parfois même, on peut utiliser les relations transversales, comme la relation d'association, pour augmenter le nombre de descripteurs concernés par une requête. Cependant, si de telles techniques[17] augmentent la quantité des résultats retournés pour une même requête, la précision de la réponse peut pâtir du manque de rigueur dans l'établissement des relations thésaurales. Les liens hiérarchiques sont en général conçus pour rendre possible un échange entre la notion subsumée et la notion subsumante lors des recherches, mais les relations transversales sont beaucoup plus floues. Ainsi, dans la branche audiovisuelle du thesaurus de l'INA, on peut observer que la notion `radio scolaire` est liée à la notion très générale d'`éducation`, qui généralise des notions qui n'ont souvent que peu à voir avec la diffusion de programmes audios, comme `cour de récréation`. L'application naïve de stratégies de reformulation hiérarchiques et transversales ne peut donc se faire sans risque important de perte de précision.

Ces faiblesses sont liées à deux facteurs que nous avons déjà évoqués :

- le manque de précision en ce qui concerne les interprétations possibles des relations de spécialisation et d'association introduites par les thesauri ;
- l'impossibilité d'introduire des relations propres au domaines, que ce soit au niveau du référentiel définissant les notions – le thesaurus – ou à celui de leur actualisation dans des descriptions de faits particuliers – les index.

On empêche ainsi la création d'informations précises, exploitables de façon pertinente et sûre par le système documentaire.

Dans le cas d'un système documentaire à base de connaissances formalisées, les spécifications contenues dans les ontologies permettent une exploitation beaucoup plus précise des index. Il est en effet possible de conduire des raisonnements qui assistent le travail du chercheur en remplissant un rôle similaire à celui des reformulations de requêtes. Mais pour cela, le système peut s'appuyer sur les significations formelles des index, obtenues à partir de celles des descripteurs qui les composent. On se place alors dans ce qu'on appelle un *cadre formel de recherche d'information*, par opposition par exemple aux approches textuelles utilisant des comparaisons entre les contenus d'index et de requêtes textuelles, aux approches numériques d'analyse et de recherche d'image reposant sur l'extraction et le calcul de similarités entre caractéristiques visuelles ou sonores du document AV, ou même encore aux approches thésaurales qui ne reposent pas sur une interprétation et des calculs formels.

Ontologies et systèmes logiques de recherche d'information

Différents modèles formels existent, qui sont employés pour fonder les processus de recherche d'information, notamment les approches numériques – telles que les calculs probabilistes de

[17] De telles généralisations peuvent être appliquées de deux manières. La première vise à compléter les index par les descripteurs qui généralisent ceux qui ont été entrés par l'indexeur, et comparer l'ensemble avec la question posée. La seconde consiste à compléter les descripteurs d'une requête par les descripteurs les spécialisant, et à comparer le résultat avec l'ensemble des index. Ces deux approches sont en théorie équivalentes quant aux résultats qu'elles permettent d'obtenir, mais en pratique les performances des systèmes sont souvent différentes, suivant les connaissances thésaurales que l'on considère, la structure et la taille des index et des requêtes, et la manière dont les complétions sont insérées dans le fonctionnement général du système. Cette distinction se rapproche de l'opposition, classique dans le domaine des systèmes experts, entre chaînage avant et chaînage arrière.

similarité entre descriptions et requêtes ([GL02] pour un exemple dans le domaine de l'audiovisuel) et les approches logiques. Celui qui guide le plus souvent les approches à base de connaissances est le modèle *logique*, inspiré par les travaux de [van86] tels que les déclinent [KC95, NB96, CMF96, Oun98, Gen00] (voir [Lal98] pour une présentation globale des différentes visions de ce modèle) : ici, on utilise l'interprétation logique formelle des index et des requêtes pour établir si les premiers sont pertinents pour les secondes. Un index i sera pertinent si un mécanisme d'inférence formelle a réussi à rapprocher « logiquement » la connaissance exprimée dans cet index de celle requise par la question q. Le plus souvent, c'est le cas si l'interprétation logique de l'index *implique* celle de la requête : $i \rightarrow q$ doit être valable pour les règles déductives[18] définies par la logique et les axiomes induits par les connaissances a priori. Car si on se place dans un cadre logique, on a néanmoins toujours affaire à un système à base de connaissances, conçu pour une application, un domaine précis. L'implication entre index et requête n'a donc pas à être prouvée *ex nihilo*, à partir des seuls axiomes de la logique : un processus de démonstration pourra utiliser des connaissances du domaine. En des termes logiques, un index pourra répondre à une requête si on parvient à démontrer que l'implication entre la représentation logique de l'index et celle de la requête est valide pour toutes les interprétations logiques satisfaisant les lois du domaine d'application. On doit avoir $T \models i \rightarrow q$, où T est un ensemble de formules – une théorie logique – traduisant lesdites lois du domaine.

C'est justement cette théorie que les ontologies sont amenées à contenir. Elles ont en effet parmi leurs buts celui de contrôler les utilisations qui sont faites du vocabulaire qu'elles introduisent, utilisation dont font évidemment partie les traitements de recherche. Elles sont à ce titre chargées d'apporter des spécifications du vocabulaire de description à même de garantir un niveau satisfaisant de cohérence interprétative. Si on considère que cette cohérence doit concerner le fonctionnement du SBC, on doit donc s'intéresser à une interprétation opérationnelle. On a déjà évoqué que des spécifications formelles étaient nécessaires à une interprétation correcte de la part d'un système informatique à base de connaissances. On l'a vu dans les exemples présentés, l'importance du paradigme logique dans les travaux d'intelligence artificielle fait que les définitions ainsi construites sont interprétées précisément selon une approche ensembliste, rapprochable de celle de la logique classique.

Dans un tel cadre, il est possible d'utiliser les spécifications ontologiques dans des processus logiques de démonstration appliqués à la recherche des index. On considère que ces spécifications sont des axiomes du domaine, des assertions toujours valides qui précisent pour un système de raisonnement logique le sens des concepts qu'il aura à manipuler. Par exemple, les spécifications données dans les codes 2.10, 2.11 et 2.12 peuvent être appliqués pour augmenter ce qui est impliqué par les assertions données en code 2.6. D'après la sémantique des logiques de description et les informations contenues dans les spécifications terminologiques, on peut en effet déduire que :

- `prof_1` joue un rôle de professeur, par application de la définition de `Professeur` ;
- `prof_1` joue un rôle d'expert, par généralisation de `RoleProfesseur`, et donc que `prof_1` est une instance d'`Expert` ;
- `emission_décrite` contient une instance d'`InterviewExpert`, par application de la définition de ce concept à l'assertion du code 2.6.

On pourra donc répondre à une « recherche d'interview d'expert », alors même que l'information n'était pas explicitement présente dans la description. Les connaissances formelles

[18]Le sens de l'implication recherchée peut varier, suivant la logique et le système d'inférence retenus. On peut se placer dans le cadre de la logique des prédicats du premier ordre classique, ou bien recourir à une logique non classique comme la logique floue. Et à l'intérieur même d'un cadre logique, les procédures de déduction validant l'implication recherchée peuvent varier : dans le cas de fragments de la logique du premier ordre, graphes conceptuels et logiques de descriptions ont recours à des stratégies de déduction différentes.

permettent d'augmenter ce qui est impliqué par l'index : on gagne en rappel par rapport à un simple mécanisme de recherche d'occurrences de notions dans les index, puisque plus d'éléments sont en mesure de répondre à une requête. Et si les spécifications sont correctes par rapport à la signification du vocabulaire dans le domaine d'application, on conserve un taux de précision élevé. En effet, les connaissances utilisées pour le raisonnement sont propres au domaine et aux usages et compréhensions qui y sont attestés, et le raisonnement lui-même repose sur des interprétations formelles précises. Le tout est de pouvoir s'appuyer sur des connaissances de raisonnements – les axiomes de l'ontologie – qui soient très précisément conçues pour garantir le respect de la continuité sémantique, qui soient en accord avec la signification des descripteurs dans le contexte applicatif de l'indexation.

Ici encore, se pose la question de la complexité des définitions formelles données dans les ontologies. Plus on donnera de contraintes sur l'interprétation du vocabulaire, plus on s'approchera des interprétations naturelles rencontrées dans l'application, et plus un système de raisonnement aura de moyens d'exploiter les index. Des études comme [SM00], ou plus directement l'observation des besoins inférentiels des systèmes à base de connaissances [GHB96, Pa03] justifient l'introduction de spécifications riches dans les ontologies, si l'application le requiert.

En particulier, il est intéressant d'encoder des *règles de production logiques* permettant de rendre compte des propriétés générales des relations, employées dans nos index (cf. figure 2.4), comme `participe`, `contient`, ou `explique`, `aPourTraitSaillant`. Ainsi, on peut vouloir rendre compte de lois comme celles de la figure 2.13 ci-dessous, extraites – comme beaucoup des exemples de ce chapitre – de l'expérimentation que nous présenterons en section 5.2.3 :

$$\forall x, y, z\ ObjetAV(x) \wedge\ ObjetAV(y) \wedge\ Anything(z) \wedge aPourTraitSaillant(x, y)\ \wedge$$
$$Explique(y, z)\ \rightarrow Explique(x, z)$$
$$\forall x, y, z\ ObjetAV(x) \wedge\ Sequence(y) \wedge\ Anything(z) \wedge Contient(x, y)\ \wedge$$
$$Explique(y, z)\ \rightarrow Explique(x, z)$$
$$\forall x, y, z\ ObjetAV(x) \wedge\ Personne(y) \wedge\ Anything(z) \wedge aPourParticipant(x, y)\ \wedge$$
$$Explique(y, z)\ \rightarrow Explique(x, z)$$
$$\forall x, y, z\ ObjetAV(x) \wedge\ ObjetAV(y) \wedge\ Personne(z) \wedge Contient(x, y)\ \wedge$$
$$aPourParticipant(y, z)\ \rightarrow aPourParticipant(x, z)$$

FIG. 2.13 – Quelques axiomes logiques pour une ontologie de l'audiovisuel

Ainsi, si un objet audiovisuel a comme trait saillant un autre objet audiovisuel – procédé de réalisation, élément de contenu visuel ou sonore, etc. – qui explique[19] un sujet donné, on pourra en déduire que le premier objet contribue lui aussi à l'explication de ce sujet. Il en va de même si par exemple un programme contient une séquence expliquant un sujet donné, ou si un objet a pour participant une personne qui explique un sujet. Finalement, on trouve dans ces exemples un axiome élémentaire de « propagation » de la relation de participation : si une personne participe à une séquence contenue dans un programme, on peut en déduire qu'elle fait également partie des participants de ce programme – elle pourra figurer dans son générique.

Avec l'index de notre exemple, on peut dès lors répondre à une requête comme « Recherche d'une émission qui explique le fonctionnement du coeur ». Le moteur d'inférence généralisera `Interview` en `Séquence`, et appliquera la règle de composition entre les relations

[19]Il est à noter qu'`explique` est une relation qui peut aussi bien s'appliquer à une personne qu'à un objet audiovisuel. On considère de fait que les deux peuvent délivrer un message explicatif... D'un point de vue formel, explique aurait donc comme domaine l'union des concepts de `Personne` et d'`ObjetAV`.

aPourTraitSaillant et explique ainsi que celle entre contient et explique. Ainsi, si l'on traduit ces règles en un langage de représentation opérationnel[20], un système pourra déduire de notre index l'ensemble des assertions de la figure 2.14.

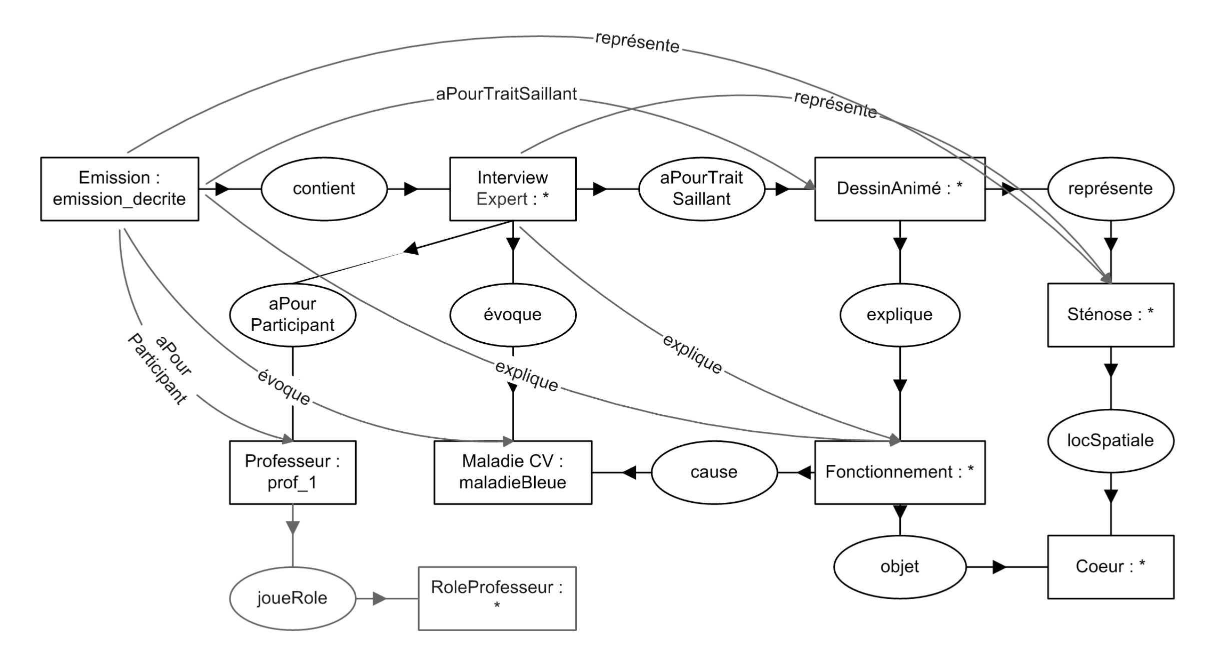


FIG. 2.14 Index complété avec des connaissances inférées

Les connaissances explicitées par le raisonnement sont représentées en gris.

Les ontologies permettent donc de mettre en œuvre des traitements qui assistent les processus de recherche de façon pertinente. En insérant les index dans un cadre d'inférence exploitant les connaissances de l'application visée dans le système, on augmente le rappel du système de recherche d'index, tout en restant précis, puisqu'on applique des stratégies de recherche qui ne sont pas indépendantes du domaine. Des SBC utilisant des ontologies autorisant la mise en œuvre de mécanismes d'inférences respectueux de la continuité sémantique peuvent donc apporter une assistance efficace aux processus d'exploitation documentaires.

2.4 Utilisations concrètes d'ontologies par des systèmes de recherche d'information

Vers des SBC répondant à nos critères fonctionnels

De fait, de nombreux systèmes d'annotation et de descriptions documentaires utilisant des ontologies plus ou moins formalisées ont déjà été conçus. Les premiers ne se sont concentrés que sur l'annotation simple de documents textuels, mais au fur et à mesure des progrès techniques et

[20]Ce qui dépend évidemment des capacités expressives du langage choisi. Les logiques de description, dans leurs formes les plus utilisées, ne peuvent par exemple pas rendre compte de toutes les règles de la figure 2.13, même si des propositions récentes ont été faites dans ce sens [GHVD03, HP04]. Les graphes conceptuels, eux, s'ils manquent par ailleurs de certaines des possibilités définitoires des LD, proposent des mécanismes d'encodage de règles relationnelles satisfaisants ([Sal97]).

de l'intérêt croissant pour les technologies du web sémantique, de nouveaux projets se sont attelés à une description et à une exploitation plus fine, en direction d'autres supports. Les observations qui suivent ne prétendent pas au statut d'état de l'art complet dans ce domaine. Elles cherchent surtout à rendre compte des problèmes rencontrés, de la manière dont ils ont pu être abordés, et en quoi les approches retenues peuvent répondre à nos préoccupations...

L'intérêt pour des techniques de représentation de connaissances structurées et spécifiées pour améliorer les processus de description et de recherche est présent depuis que les chercheurs s'efforcent de trouver une alternative aux systèmes reposant sur la recherche textuelle. Par exemple, [AS92] proposait déjà de prendre en compte l'importance des relations entre entités dans un modèle de données graphique, et un système de recherche reconnaissant les parcours possibles dans ces graphes pour répondre à des requêtes. Mais le référentiel conceptuel employé était relativement limité : quelques *types* seulement, et non organisés entre eux. Au fur et à mesure des efforts de recherche, ce référentiel s'est progressivement étoffé. Certains travaux d'inspiration plus linguistique ont également constaté de leur côté l'importance de la structuration relationnelle des données à exploiter. Le projet MENELAS [ZC94] a recours à des référentiels conceptuels et relationnels riches pour guider l'extraction d'information structurées à partir de textes, en faisant le pari que ces référentiels – créés suivant les propositions méthodologiques que nous verrons en chapitre 3 – feront bénéficier les résultats du système d'une certaine forme de continuité sémantique. Le projet CONCERTO [ZBB+99], que nous reverrons lui aussi plus tard, s'efforce également de concevoir une ontologie – élaborée suivant des critères d'analyse sémantique de textes – pour améliorer description et recherche. Cependant, les processus de recherche de ces systèmes n'exploitent pas encore des connaissances de raisonnement aussi riches que celles que nous recherchons. Ainsi, le système SYDOM [RCP02], utilise des graphes conceptuels adaptés à la représentation et la recherche multilingue, insiste bien sur l'importance de la définition des concepts, mais les mécanismes de recherche – basés sur la similarité entre graphes – sont relativement génériques, dans la mesure où ils n'exploitent pas les propriétés de relations particulières.

A l'inverse, des travaux comme [KC95] se concentrent principalement sur la mise en place de capacités « logiques » pour des raisonnements complexes répondant aux besoins observés. David Genest, lui aussi, [Gen00] propose de suivre des stratégies de raisonnement prenant réellement en compte les besoins d'un domaine applicatif et exploitant les relations entre concepts. Le problème est que ces travaux se concentrent plutôt sur la création de solutions techniques[21], mais ne se placent pas dans un cadre méthodologique complet traitant à la fois de l'utilisation et de la conception du contenu des ressources conceptuelles construites.

Ces travaux, de plus, n'abordent cependant pas explicitement les documents audiovisuels. Dans le domaine du traitement logique de représentation de l'image – fixe – les travaux du laboratoire MRIM font référence. [MBC95] introduit un modèle de description spécifique à l'image, reposant sur plusieurs point de vues – structurel, spatial, perceptif, symbolique – qui sont censés en épuiser les différentes dimensions. Le contenu des descriptions lui-même est représenté à l'aide de graphes conceptuels, comme dans [OP98] et [MCMO03]. Si les mécanismes de recherche proposés tiennent bien compte des propriétés du medium visuel lors de la description et de la recherche, l'adaptation de ces mécanismes aux entités des domaines particuliers qui constituent le contenu thématique des documents semblent cependant plus problématique. Il faut définir une approche plus générique, où les ressources conceptuelles seraient modifiables suivant les applications rencontrées.

[21] Ce qui est d'ailleurs normal, pour des travaux qui ont été menés à une époque où de telles solutions n'existaient pas. Notre travail n'a sans aucun doute pu se positionner de manière différente que parce que nous pouvions justement déjà utiliser des outils aussi intéressants que COGITANT [GS98].

A cette préoccupation, l'initiative du *web sémantique* peut apporter une réponse. Récemment, l'un des initiateurs du Web, Tim Berners-Lee, a développé avec d'autres chercheurs la vision d'une extension du web actuel, où l'information serait pourvue d'une signification plus précise que celle d'aujourd'hui, qui demeure peu ou prou inaccessible aux systèmes informatiques. Les contenus des documents seraient conçus comme autant de ressources disponibles pour un réseau d'agents logiciels échangeant des services les uns avec les autres, afin de permettre un meilleur accès à l'information pour les utilisateurs humains [BHL01].

Il faut pour cela étendre les formats actuels de contenu, destinés à une lecture humaine, par une couche « sémantique » apportant des méta-données ou des annotations interprétables par un système en vue de la conduite de raisonnements automatiques. Il faut également permettre l'établissement de liens entre les ressources sémantiques possiblement hétérogènes pour mieux les échanger et les intégrer. On doit d'abord résoudre un problème d'*interopérabilité syntaxique* : les données doivent être encodées dans un format compréhensible par tous les agents. XML [XML04] et le langage générique de description de ressources RDF[22] [RDF04a], tous deux développés par le W3C, fournissent un cadre expressif standard, et apportent une première réponse. Mais subsiste un problème d'*interopérabilité sémantique* : les ressources publiées sur le web peuvent correspondre à des applications et des conceptualisations différentes : il n'est pas réaliste d'imposer un seul schéma conceptuel à l'échelle du web. Il importe donc :

- d'expliciter les conceptualisations dont relèvent les ressources ;
- de spécifer la manière dont des conceptualisations différentes peuvent être articulées.

Pour tout cela, les ontologies peuvent apporter une aide précieuse. Comme on vient de le voir dans ce chapitre, elles constituent des référentiels de sens pouvant améliorer les communications entre agents, tout en garantissant un degré de pertinence suffisant pour les traitements qui seront effectués de manière automatique à partir des éléments de spécification qu'elles fournissent. Dans la vision du web sémantique, les ontologies permettent le contrôle et l'exploitation de méta-données s'insérant dans un réseau de ressources structurées. Ces données constituent alors des composants de connaissance partageables et exploitables « intelligemment » par des SBC pour la recherche d'information, l'aide à la navigation, etc. Et comme les domaines potentiellement modélisables sont infinis, les efforts de recherche doivent se concentrer sur la création d'outils génériques acceptant autant de conceptualisations que d'applications et domaines considérés. Ainsi que sur la réflexion méthodologique dans le but de pouvoir gérer de la manière la plus efficace possible la conception et l'utilisation de ces référentiels spécialisés...

Cette initiative du web sémantique a généré quelques systèmes de recherche qui sont devenus par la suite des références. **SHOE** [HH01] et **Ontobroker** [FAD+00] par exemple, ont été les premiers à gérer la représentation d'annotation documentaires reposant sur des ontologies qui permettent l'adaptation des traitements dont ces représentations font l'objet. Projets pionniers, ils ont cependant plus cherché à répondre aux problèmes techniques qu'à se positionner du côté des usages tels que nous les concevons. **Ontoseek** [GMV99] se rapprochait plus de cas d'utilisation réels, et des utilisateurs que l'on peut y rencontrer – par l'utilisation de l'ontologie linguistique WORDNET [Fel98]. Mais les procédés de recherche, utilisant la similarité entre graphes, ne répondent plus forcément à ce que nous cherchons. Quant aux travaux exposés dans [RBF+02, AMO+03, HS03a, Svd+04], qui représentent autant d'avancées importantes pour la réalisation technique des objectifs du web sémantique, il n'opèrent pas vraiment de rapprochement explicite et détaillé entre ces solutions et les besoins auxquels elles permettent de répondre, rapprochement indispensable à l'obtention d'un cadre méthodologique cohérent. De plus, ces projets n'abordent pas de thème semblable à celui qui a motivé cette thèse, celui de la produc-

[22] On verra en section 5.2.3 un exemple d'utilisation de ce langage.

tion d'index riches dans le cadre de documents av, index dont on espère retirer le plus possible pendant le fonctionnement du système de recherche.

En ce qui concerne l'aspect audiovisuel, certains projets visant l'amélioration de l'accès ou de l'exploitation de documents *multimedia* apportent des contributions intéressantes. Par exemple, [GHB96], s'il ne se place pas dans un cadre méthodologique global, procède à une analyse des besoins en matière d'expressivité et de raisonnement qui se rapproche de notre point de vue. [KHPG02, MAH03, DDH+04] proposent eux des solutions techniques adaptées à l'annotation de documents en tenant compte de leurs spécificités (sélection de régions spatiales, gestion de la dimension temporelle) mais ils n'offrent pas de cadre de recherche exploitant de manière satisfaisante les possibilités de raisonnement offertes par des ontologies complexes. [LH04] présente une manière originale de créer des connaissances de raisonnement riches et utiles pour l'exploitation de données dans un cadre applicatif concret – la reconnaissance automatique de concepts visuels de « haut niveau »[23] – mais les outils proposés ne sortent pas de ce domaine applicatif précis.

Les projets MuseumFinland [HSV04] et MIA [SDWW01, SBC+02, HSWW03] se placent eux dans un contexte qui est beaucoup plus proche. Ils cherchent en effet à créer et exploiter des descriptions utilisant des descripteurs issus de thesauri dédiés à des domaines multiples dans des systèmes reposant sur les solutions du web sémantique, réellement adaptables à des applications précises, et visant la description et la recherche de contenus thématiques qui peuvent être ceux de documents visuels. Par exemple, [HSV04] propose des « règles de recommandation » qui constituent de véritables connaissances de raisonnement permettant de relier des objets entre eux et d'améliorer la recherche dans des bases documentaires volumineuses. Malheureusement, ce travail souffre du point de vue de l'expressivité de la limitation engendrée par le contexte thésaural initial : en dehors des schémas de descriptions utilisés de manière standard dans les musées, il est très difficile d'obtenir des structures relationnelles aussi précises que celles que nous désirons. Nous verrons plus en détails dans la section 3.3.1 (p. 3.3.1) les aspects qui nous séparent des travaux néerlandais.

Enfin, il faut mentionner les travaux de Jane Hunter [Hun03] et de Chrisa Tsinaraki [TPC04] qui si elles se concentrent surtout sur les problèmes d'interopérabilité des descriptions – parfois de bas niveau – et non sur le raisonnement et la recherche, proposent tout de même des ressources ontologiques pouvant convenir au cas des documents audiovisuels, dont on peut s'inspirer pour la création de référentiels exclusivement dédiés à la modélisation des documents AV, comme celui présenté dans la section 5.2.3 de ce manuscrit.

Il existe donc une grande variété d'initiatives de recherche cherchant à utiliser les ontologies dans un cadre documentaire. De fait, la multiplication de ces projets a conduit à fixer des paradigmes et des méthodes bénéficiant d'un intérêt qui dépasse – et de loin – le cadre de laboratoires isolés. L'initiative du web sémantique, en particulier, a fédéré une communauté de vues et de moyens tout à fait propice à la maturation accélérée des techniques proposées.

Néanmoins l'indexation audiovisuelle dans le cadre d'un domaine interprétatif donné n'est pas le contexte applicatif affiché par le plus grand nombre, loin s'en faut. Et toutes les approches évoquées ne s'attachent pas explicitement sur les objectifs que nous avons isolés dans le chapitre 1. Il est donc pertinent pour l'INA de poursuivre un effort de recherche allant dans cette direction. On peut d'ailleurs noter que ces conditions, valables au moment où le projet Opales a été lancé, le sont encore au moment de la rédaction de ce mémoire. Peu d'études ont cherché à la fois à prendre en compte explicitement les usages de l'indexation audiovisuelle – et les objectifs qu'ils

[23] C'est-à-dire demandant une forme d'interprétation qui va au-delà de l'analyse des composantes physiques d'une image.

imposent – dans le cadre technique et méthodologique offert par les approches de représentation de connaissances ontologiques.

Le cas d'Opales C'est effectivement une approche tout à fait orientée vers les usages qui a été adoptée par Opales. Il s'agit en effet en premier lieu de trouver une solution au problème d'une indexation dont les termes et les enjeux varient suivant le *point de vue* applicatif retenu : les deux expérimentations qui se sont déroulées lors du projet concernaient la petite enfance – point de vue mêlant des considérations sociologiques, ethnologiques et psychologiques – et l'analyse pédagogique de documentaires géographiques pour mettre en lumière l'articulation entre procédés audiovisuels et présentation du discours scientifique. Pour cela, la plate-forme d'indexation doit proposer la création d'un vocabulaire spécifique à chaque point de vue et des modes de création et d'exploitation des index qui répondent à ce que l'on attend dans chaque application.

Les outils réalisés permettent donc la conception d'ontologies dont la partie formelle est exprimée sous la forme de supports de graphes conceptuels (cf. section 4.3.3), format dont certains partenaires du projet étaient spécialistes. Ces ontologies sont ensuite proposées à l'utilisateur pour créer des index des segments documentaires qu'il a isolés (cf. figure 2.15). Le système se charge de vérifier la cohérence des descriptions par rapport aux spécifications formelles – signature des relations et liens hiérarchiques – et applique des connaissances de raisonnement spécifiques à chacune des ontologies – données sous la forme de règles – pour améliorer les résultats des recherches. Pour gérer les annotations conceptuelles et conduire les inférences, Opales utilise Cogitant [24] [GS98], développé par le laboratoire lirmm-gc.

Du point de vue du parcours succinct que nous venons de faire, Opales doit donc envisager des problèmes généraux de représentation et de gestion des connaissances, dans un cadre logique où les index ont une interprétation formelle mobilisable en vue de l'assistance à la recherche par le raisonnement logique. Tout cela doit être traité avec le souci constant de la spécificité du cas audiovisuel ; il faut notamment trouver des solutions qui respectent voire facilitent la continuité sémantique, comme le contrôle automatique du processus d'indexation. On a donc affaire à un large éventail de problèmes, qui ne sont que partiellement communs avec ceux du web sémantique, puisqu'ils insistent sur des points qui sont parfois laissés de côté dans cette initiative.

2.5 Conclusion

Nous avons montré qu'il était envisageable de trouver une solution aux trois problèmes fonctionnels dégagés dans le chapitre précédent. On peut effectivement créer des systèmes documentaires où les index et les traitements qui leur sont appliqués seront spécifiés au niveau de la connaissance. De fait, les techniques de représentation des connaissances permettent de créer des descriptions structurées : des langages de représentation mobilisant à la fois concepts et relations sont utilisés pour rendre compte de la complexité du contenu documentaire. Comme dans les systèmes documentaires classiques, il est possible et même souhaitable de munir de tels langages de représentation d'un vocabulaire dédié à l'application que l'on envisage. Dans le cadre des sbc, on a vu que ces vocabulaires sont créés dans ce que l'on appelle des ontologies. A la manière des thesauri, celles-ci introduisent les notions servant à la description tout en guidant leur interprétation. On peut alors envisager l'instauration d'une continuité sémantique dans les processus documentaires, et cela à deux titres.

Tout d'abord, nous avons constaté que les ontologies sont des spécifications plus contraintes que ne le sont les thesauri. Les méta-langages qui permettent leur création utilisent en effet des

[24] `http://cogitant.sourceforge.net`

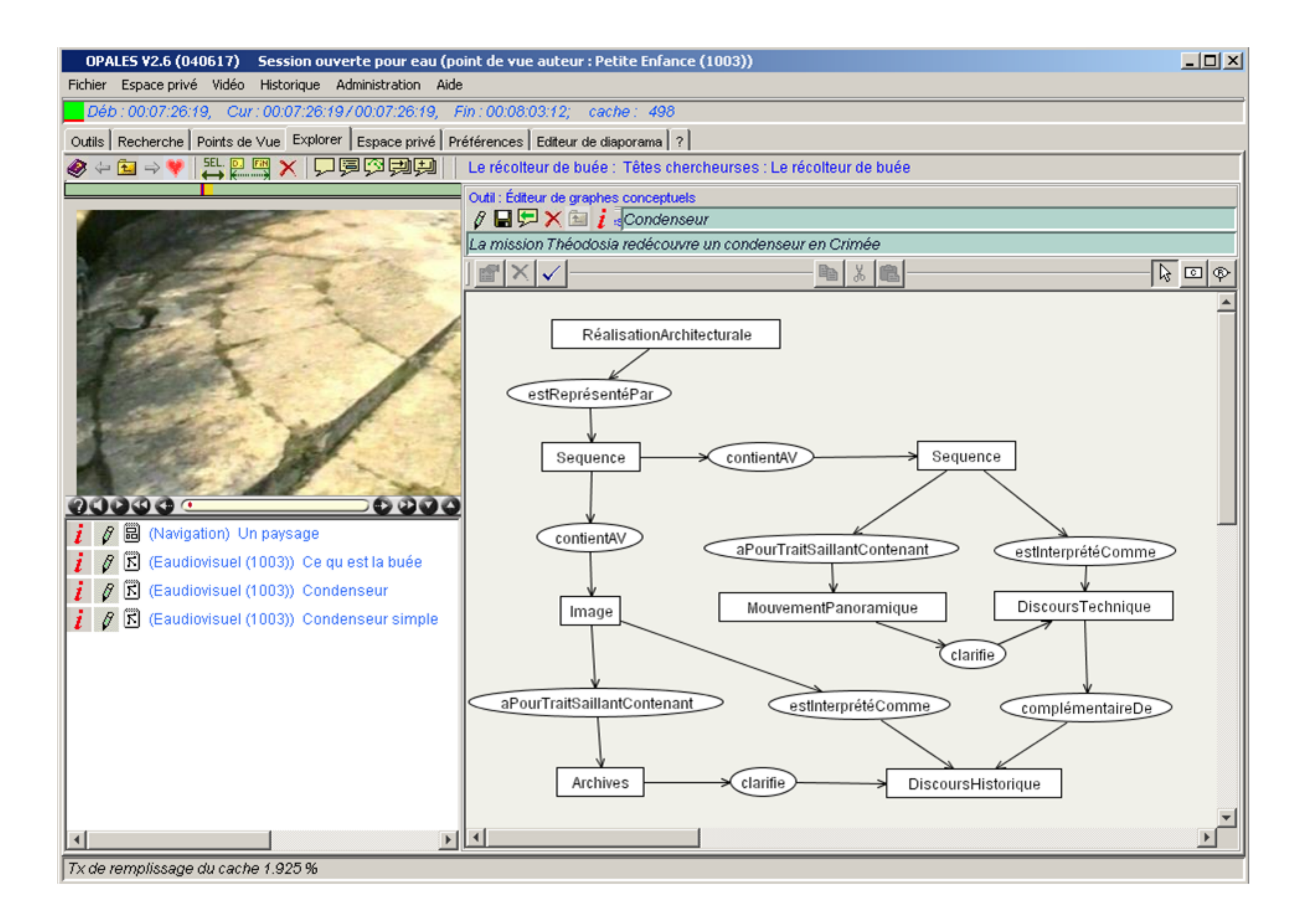


Fig. 2.15 Un gc index dans Opales

primitives de représentation reposant sur des interprétations formelles. Par rapport aux systèmes documentaires existants, il y a là une grande avancée. On a en effet vu que la signification des liens des thesaurus est trop floue pour garantir la véracité d'interprétations opérationnelles qui seraient dérivées directement du réseau de relations sémantiques. Et même si des traitements d'inspiration « graphique » exploitant les distances entre notions peuvent être mis en œuvre dans des thesaurus adaptés à des domaines applicatifs extrêmement précis ([CTHD00]), on applique toujours des stratégies de recherche et de reformulation qui sont génériques, indépendantes du domaine. En un sens, l'engagement ontologique de telles solutions n'est pas assez fort pour autoriser une exploitation plus fine.

Les ontologies, reposant sur des hiérarchies de spécialisation exploitables de façon comparable aux thesauri, semblent dans un premier temps apporter des solutions similaires. Cependant, leurs concepts et leurs relations bénéficient d'un engagement ontologique différent, favorable à une plus grande précision des systèmes de manipulation des descriptions. Et assez rapidement, les solutions ontologiques apparaissent comme les seules pouvant autoriser des stratégies fines de recherche utilisant les liens spécifiques du domaine. Il devient possible de créer des spécifications relationnelles plus riches que celles des thesauri, ainsi que d'exploiter ces spécifications de manière pertinente pour rapprocher le travail de l'indexeur de celui du chercheur, dans le cadre d'une interprétation sémantique partagée du vocabulaire d'indexation. Le système prend alors en charge lui-même une partie des efforts nécessaires à la conservation de la continuité sémantique. Cela est valable tant au niveau du contrôle de cette continuité lors de la création des index que lors de l'application de processus d'inférence qui utilisent les spécification ontologiques pour rapprocher les index créés des requêtes adressées au système.

Si certaines des connaissances ainsi explicitées peuvent sembler triviales, il ne faut pas oublier que ce sont probablement celles-ci qui sont jugées intéressantes par l'utilisateur, puisqu'il n'a plus à opérer les traitements de reformulation les plus évidents. Par exemple, dans un cadre de recherche documentaire utilisant un thesaurus, [GS93] montrait comment l'utilisation de systèmes experts exploitant les hiérarchies diminue le travail de celui qui recherche une information, qui peut par conséquent concentrer son activité sur d'autres points, plus subtils, de la recherche documentaire. Le recours à des systèmes ontologiques va encore plus loin, puisque c'est justement la prise en compte efficace des relations propres au domaine d'application, qui est reconnue comme un point d'amélioration des systèmes de recherche documentaire existants [TAJ01, VPS03, Blo04].

La recherche concernant les ontologies et les SBC qu'elles permettent de créer constitue un domaine très actif. De nombreux projets ont vu le jour, notamment dans le contexte du web sémantique. Pour l'INA, l'enjeu est quelque peu différent : s'il est indispensable de maîtriser les propositions techniques qui sont amenées à émerger de ces initiatives, il faut faire en sorte de les adapter à la problématique de l'indexation des documents audiovisuels et à l'utilisation d'une telle représentation de leur contenu dans les systèmes de recherche documentaires.

Notre thèse a été initiée dans cette optique, et a cherché à retirer des expérimentations réalisées dans le projet OPALES ou dans les initiatives proches un cadre cohérent et pertinent de conception et d'utilisation des ressources ontologiques dans les processus documentaires. De fait, beaucoup d'études, et c'est légitime dans un champ de recherche en cours d'élaboration, se sont concentrées sur les difficultés d'ordre plutôt technique : codage des annotations qui servent de support à l'indexation, représentation opérationnalisable des connaissances de raisonnement, mise en œuvre de processus d'inférence exploitant ces raisonnements... Nous cherchons plutôt à envisager ces problèmes techniques non en eux-mêmes, mais en tant que moyens mis au service de

scénarios d'utilisation réels. Comment une approche d'indexation à base de connaissances peut être correctement mise en place ? Peut-on cacher la complexité inhérente à une telle approche aux yeux de l'utilisateur humain ?

Une réflexion globale comme celle de [MSD00], récapitulant les enjeux de l'utilisation des ontologies dans les systèmes documentaires et les confrontant aux projets expérimentaux effectivement conduits, évoque bien la difficulté que nous avons partiellement mise au jour. Et [WSWS01], qui a essayé de transposer les connaissances contenues dans un thesaurus à un SBC ontologique, montre que le passage d'une approche classique à une vision plus formelle et plus riche ne peut pas faire l'impasse sur des réflexions méthodologiques indispensables à l'efficacité du système. Par exemple, l'utilisation des ontologies est tout à fait justifiée par les possibilités d'assistance permises par les moteurs d'inférence qui les exploitent. Il faut dès lors que les spécifications formelles soient clairement orientées vers l'utilisateur, que les raisonnements qu'elles permettent soient effectivement pertinents pour les usages retenus. Si légitimer le recours aux SBC est tout à fait faisable, encore faut-il fournir des éléments qui facilitent l'application pratique des approches évoquées au long de ce chapitre. Que ce soit pour la maîtrise d'une expressivité désormais plus importante, pour la conception et l'appréhension de référentiels de signification qui sont très précis et dont certains des aspects formels ne sont pas immédiatement accessibles pour le non-expert, ou encore en ce qui concerne la proposition et la mise en œuvre de connaissances de raisonnement effectivement pertinentes, il faut des considérations méthodologiques à même de guider le travail des concepteurs et des utilisateurs du système.

Deuxième partie

Prendre en compte l'usage dans l'implémentation de solutions d'indexation ontologique

Chapitre 3

Faciliter la conception et l'accès aux index sémantiques dans un système documentaire

3.1 Introduction : vers une prise en compte des usages dans les solutions existantes

Recourir à une approche sémantique employant des ontologies complexifie le processus d'indexation et d'accès aux index. L'utilisateur se retrouve en effet directement confronté au référentiel conceptuel et formel de représentation des connaissances lors de deux étapes cruciales pour le système documentaire :

- quand il doit l'utiliser pour produire un index qui soit pertinent ;
- quand il doit, le plus souvent en tant que chercheur, accéder à la signification d'un index déjà existant, renvoyé par le système lors d'une requête, par exemple.

Comment faire alors pour faciliter la tâche de l'utilisateur à ces deux moments ? De la même manière que pour les SBC se pose, d'un point de vue technique, le problème d'un compromis entre expressivité autorisée et traitements envisageables d'un point de vue opérationnel, on doit aboutir à un équilibre satisfaisant entre la complexité d'un système de gestion des connaissances d'une part, et son utilisabilité par des agents humains d'autre part. Cette utilisabilité sera évidemment évaluée à l'aune des pratiques déjà existantes, et des capacités cognitives des utilisateurs, qui ne maîtrisent pas forcément les ressorts de la sémantique formelle ou opérationnelle des ontologies. Les ontologies doivent donc apporter des spécifications qui permettent d'obtenir à la fois :

- une indexation économe d'un point de vue cognitif, et des processus de recherche naturellement compréhensibles ;
- des traitements formels efficaces, qui s'appuient sur des significations formelles élaborées pour garantir la cohérence et la précision qui ont nécessité dans les systèmes documentaires l'abandon de la langue non contrainte.

Ces deux objectifs apparemment contradictoires peuvent, on va le voir, être partiellement réconciliés. On peut pour cela s'appuyer sur les pratiques existant dans le domaine d'application du système documentaire : les éléments qui sont intéressants à décrire, leur interprétation dans le cadre applicatif, et l'assistance attendue de la part du système de recherche. On l'a vu en effet dans le chapitre 1, on n'indexe le plus souvent que dans le cas d'une application bien précise. Sans visée applicative, pas de stratégie d'interprétation du contenu documentaire, et pas – ou très peu – de prescriptions interprétatives à même de garantir une compréhension et une ex-

ploitation correctes des index. Ce constat, évident dans le cadre documentaire traditionnel, est souvent oublié ou négligé dans les approches à base de connaissances.

Est-il faisable d'isoler de manière précise les pratiques applicatives ? Il n'est pas du ressort de ce manuscrit d'apporter une réponse définitive à ce problème général d'acquisition des connaissances. Néanmoins, on peut faire remarquer que les usages documentaires s'incrivent souvent dans des communautés bien délimitées (recherche scientifique dans un domaine donné, gestion du fonds documentaire d'une entreprise). Et que, comme c'était le cas dans le projet Opales, un tel contexte est propice au figement d'un certain nombre de comportements et de compréhensions. Des pratiques pertinentes sont donc identifiables, tant en ce qui concerne la description que la recherche documentaires. On connaît le vocabulaire de l'application, et on peut cerner les besoins d'indexation qui y sont liés. Même dans le cadre de l'INA, qui indexe des documents relevant d'une très grande diversité de thèmes, on peut isoler des pratiques d'indexation récurrentes, liées à la spécificité du document audiovisuel et des exploitations ultérieures que l'on peut envisager. On s'attachera par exemple à la description des procédés audiovisuels, des dispositifs employés pour présenter un thème ...

Même si l'élicitation des usages applicatifs est délicate – ce qui revient à affirmer que l'usage n'est pas clairement identifiable – on peut toujours se reposer sur l'adoption de bonnes pratiques qui confèrent aux ressources ontologiques le statut de système prescrivant une interprétation et des usages dans un cadre restreint par l'outil documentaire. En fait, on peut affirmer que tout système à base de connaissances doit obligatoirement réaliser une forme d'auto-prescription de son usage, puisqu'aucune application courante n'exploite de manière naturelle de connaissances formalisées. Et qu'en fait la vision d'un sbc isolé du monde ne devant compter que sur ses propres ressources pour guider son utilisation n'est pas si absurde que cela. Mais nous préférons laisser là cette discussion, et considérer qu'il est toujours envisageable d'avoir accès à des usages applicatifs extérieurs au système, et donc de réfléchir à une assistance conforme à ce qui est attendu dans ces applications.

Là encore, il s'agit de respecter une certaine forme de continuité sémantique : le passage par les ressources ontologiques ne doit pas créer de rupture par rapport aux usages. Si les informations contenues dans les index et les traitements que le système opère sur ceux-ci font consensus, alors le système aura la crédibilité qui est nécessaire à son adoption. Finalement, on demande aux ontologies de respecter ce pourquoi elles apparaissent si centrales aux systèmes, au-delà du fait de servir de simple catalogue d'éléments d'indexation. Partage et consensus sont en effet des notions qui reviennent souvent dans les définitions proposées. Mais la nécessité de l'articulation entre le niveau des connaissances du domaine, souvent exprimées à l'aide de la langue *naturelle*, et le niveau formel, débouchant sur les spécifications opérationnelles des calculs autorisés dans le système, est loin d'être systématiquement prise en compte.

Nous allons exposer dans ce chapitre les méthodes proposées dans les projets existants, et comment ces méthodes sont mises en œuvre par des outils concrets, ce qui nous poussera à réexaminer les solutions déjà présentées, en gardant cette fois-ci à l'esprit les trois points méthodologiques dégagés dans l'introduction de cette partie. Puis nous présenterons les solutions que nous avons avancées ou bien reprises et appliquées méthodiquement dans le cadre de notre travail de thèse. Nous nous attacherons en premier lieu aux moyens de rendre les termes de l'indexation compréhensibles, en voyant comment on peut les relier aux compréhensions rencontrées dans le domaine applicatif. Nous montrerons ensuite qu'il est possible d'assister la création des index eux-mêmes par des préconisations d'indexation. Et que l'on peut concevoir des connaissances

de raisonnement aisément reconnues comme pertinentes, puisqu'elles vont contribuer à abaisser les contraintes imposées à l'indexeur ou au chercheur tout en laissant au système des capacités intéressantes en ce qui concerne l'appariement entre index et requêtes.

3.2 Aider la compréhension de la substance des descriptions

3.2.1 Le nécessaire ancrage dans les compréhensions et usages du domaine

Il ne faut donc pas oublier qu'un système documentaire doit permettre au document indexé d'être utilisé dans des pratiques précises. On a besoin d'une information qui soit exploitable pour un usage donné, ce qui suppose, dans le cas de l'indexation, le recours à un langage adapté et des procédés de traitement pertinents s'appliquant aux expressions de ce langages (cf. chapitre 1).

Dans un système utilisant la langue, celui qui indexe ou qui accède à un index est placé dans un contexte interprétatif fort, qui fournit aux termes employés pour la description des significations accessibles, sinon évidentes. Le partage de ces interprétations au sein de la communauté est alors le gage d'une certaine forme de continuité sémantique.

Dans le cadre des systèmes employant des thesauri, la situation devient plus complexe. Le vocabulaire est contrôlé, les descriptions n'apparaissent plus dans des textes – écrits ou oraux – susceptibles de favoriser une interprétation conforme à ce qui a lieu dans la communauté d'usage concernée. Pour cela, on l'a vu dans le chapitre 1, le thesaurus essaie de se rapprocher au maximum de son contexte d'usage.

Tout d'abord, il s'efforce d'associer aux descripteurs qu'il introduit des signifiants tirés de la langue telle qu'elle est pratiquée dans le domaine d'application. Employer des mots permet en effet de bénéficier de la signification qu'ils obtiennent en temps normal : le terme `best of`, présent dans l'extrait du thesaurus de l'INA de la table 3.1, appartient clairement au vocabulaire spécialisé de l'analyse du document audiovisuel. Si on avait voulu rendre compte plus clairement de la signification de cette notion hors de son contexte, on aurait sûrement choisi une autre étiquette, comme « florilège », qui est reconnu comme terme proche mais ne peut être employé comme descripteur. Mais on se serait alors coupé de la pratique qui mobilise le descripteur, ce qui lui aurait retiré de son intelligibilité et, par conséquent, de son utilisabilité.

Ensuite, les thesauri couplent ce choix d'intitulés langagiers pertinents avec une stratégie complémentaire de re-création d'un contexte d'interprétation. On a vu comment les relations sémantiques (spécialisation, synonymie, association) précisent l'interprétation d'une notion. Pour achever le travail d'explicitation des significations de ces notions, beaucoup de thesauri ont recours à des notes d'application, que nous avons déjà introduites en page 25. Ces brèves annotations textuelles donnent à celui qui accède au thesaurus des instructions sur la manière d'employer le descripteur, ou bien lèvent une ambiguité qui pourrait subsister quant aux rapports entre le descripteur et un autre concept, d'acception ou de formulation proche.

Par exemple, le thesaurus de l'INA, dans sa branche sur les formes artistiques, comporte une grande quantité de ces notes pour indiquer aux documentalistes des subtilités d'emploi des descripteurs, comme le fait que l'`art brut` fait référence à des œuvres conçues indépendamment d'une référence à un courant, et que ce descripteur ne doit être employé que si l'œuvre est postérieure à 1945. De même, dans la branche audiovisuelle, la note d'application de `best of` désigne un produit obtenu à partir d'extrait d'une même collection, ce qui l'oppose – doublement ! – à `bêtisier` : l'absence de note d'application pour celui-ci laisse entendre que la provenance de ses éléments est sans importance. On peut finalement mentionner l'exemple du descripteur `remontage`, intéressant à double titre :

– on a affaire à un terme spécialisé, qui est interprétable correctement du fait de sa situation dans la portée de `audiovisuel` : le même terme sous une notion liée à la mécanique aurait une interprétation tout à fait différente ;
– le fait que cette notion soit placée sous un type de `produit`, conjugué à la note d'application, lève une ambiguité qui subsisterait même dans le domaine de l'audiovisuel, puisque le terme peut dans ce domaine désigner à la fois une action et le résultat de cette action.

```
époque artistique
        art contemporain
                art brut
                NA :Productions nées spontanément sans références
                aux écoles, courants...
                NH :A partir de 1945
...
audiovisuel
        produit audiovisuel
                best of
                NA :Morceaux choisis d'une collection
                UP :florilège
                bétisier
                ...
                remontage
                NA :Forme nouvelle d'un produit audiovisuel déjà diffusé
```

Tab. 3.1 Notes d'usage dans le thesaurus de l'INA

`NA` : note d'application, `NH` : note historique

En sus de ces informations internes au thesaurus, on peut constater que ces constructions sont souvent accompagnées de guides extrêmement complets sur les descripteurs qu'il contient. Michel Dauzats, analysant quelques thesauri du domaine de l'audiovisuel, confirme que la plupart viennent avec une documentation volumineuse [Dau94].

De fait, les thesauri sont reconnus par les spécialistes de la documentation comme des objets complexes, qui demandent une assez longue période de familiarisation avant de pouvoir être utilisés correctement dans les processus de description et de recherche. Et pourtant, même si ces spécialistes rechignent à laisser les non-experts seuls face aux systèmes thésauraux, force est de reconnaître que les thesauri apportent des spécifications qui les mettent à portée d'un public plus large. Au moins peuvent-ils guider les interprétations de ceux qui ont des notions basiques dans le domaine qu'ils concernent. . .

Le cas des ontologies est encore plus délicat. Il est évident qu'on ne peut plus se passer d'une assistance à la compréhension, l'ontologie étant un objet comportant des spécifications formelles très éloignées des significations « métier ». Les ontologies doivent donc bénéficier d'efforts qui poursuivent ce qui est entrepris pour les thesauri.

Le premier réflexe est le recours aux étiquettes d'essence linguistique. Comme dans les thesauri, les concepts et les relations d'une ontologie sont généralement exprimés à l'aide des termes du domaine visé, comme dans les figures 2.7 et 2.8 des pages 55 et 56. Cela apparaît de manière évidente dans le cas des projets qui s'intéressent à la représentation des connaissances « multi-

lingue », comme [RCP02]. Ici, concepts et termes sont distingués, mais explicitement reliés les uns aux autres. Les relations, définies au niveau conceptuel par leur signature, sont elles aussi reliées à des intitulés linguistiques. On obtient donc une intelligibilité minimale pour les concepts et les relations d'une ontologie. Mais il ne s'agit là que d'une « bonne pratique » dont le résultat n'est pas acquis. En dernier ressort, tout dépend de la capacité du concepteur de l'ontologie à choisir des symboles linguistiques appropriés, ce qui est loin d'être évident en l'absence de cadre méthodologique précis, comme c'est la plupart des fois le cas. Pour gagner en crédibilité, des outils ([CM01, KV03], par exemple) essaient de relier les concepts d'une ontologie à des terminologies reconnues, telles que WordNet [Fel98]. Mais si ces terminologies ne sont pas obligatoirement conçues avec un besoin spécifique en ligne de mire, on ne peut garantir la pertinence des résultats obtenus en termes d'intelligibilité dans des cas applicatifs précis. Pour des terminologies liées à un domaine précis – SNOMED dans le cas de [CM01] – l'intérêt est plus évident, mais la donnée de tous ces termes ne prescrit une interprétation précise que si celui qui accède à une telle ontologie maîtrise déjà le référentiel terminologique utilisé.

Texte	N° de la phrase - retour au corpus	UL 1	UL 2	Enoncé définitoire	Relation sémantique entre UL1 et UL2
CorpusPE-FS	495	présente des plaques d' urticaire	eczema	Soyons clair , la peau de votre enfant , sauf si elle présente des plaques d' urticaire (eczema) , n' a besoin que :	UL2 est hyperonyme de (est hyperonyme de / est paradigme de / est synonyme de / est en rel. fonct. avec)
CorpusPE-FS	647	Bureau of educational research	BER	Les Whiting ont établi à l' Université de Nairobi , au Kenya , un institut de recherche (devenu Bureau of educational research (BER) qui a permis de former une quantité de chercheurs , aussi bien africains qu' américains .	est hyperonyme de UL1 OK
CorpusPE-FS	726	nécessitent une approche ethnologique et de	inter	Tous ces cas nécessitent une approche ethnologique et de psychologie (inter) - culturelle .	UL2 est hyperonyme de UL1 OK
CorpusPE-FS	927	bien faire	faire ce que la mère veut	C' est à cette époque que s' enracine dans l' individu la morale du bien faire (faire ce que la mère veut) voire du trop bien faire si la pression parentale est trop forte .	UL2 est hyperonyme de UL1 OK
CorpusPE-FS	1220	mangent davantage de fruits	agrumes	En l' occurrence , les femmes enceintes mangent davantage de fruits (agrumes) et des légumes (soja) classés comme froids afin que leur lait , perçu comme chaud , soit suffisamment bon et abondant pour l' enfant qui va naître .	UL2 est hyperonyme de UL1 OK
CorpusPE-		et des légumes	soja	En l' occurrence , les femmes enceintes mangent davantage de fruits (agrumes) et des légumes	UL2

Fig. 3.1 Extraction de termes et d'énoncés définitoires, extrait de [MZB04]

Une ligne correspond à une détection réussie d'un patron lexico-syntaxique, elle présente les deux unités lexicales liées, et le type du lien sémantique détecté.

Certaines approches proposent de rattacher automatiquement les notions ontologiques à des

textes du domaine qui leur fourniraient un contexte d'interprétation satisfaisant. De fait, cette vision est liée à tout un courant préconisant la construction d'ontologies à partir de ces textes mêmes (voir [BAC04b] pour un aperçu récent). Non seulement des outils d'analyse textuelle extraient des termes, mais ils s'efforcent de les rattacher en permanence, pendant le processus de construction puis dans l'ontologie construite, aux passages textuels desquels ils ont été extraits. Certains vont plus loin en exploitant les énoncés particuliers qui ont une valeur définitoire pour le terme extrait. En repérant des marqueurs lexicaux, employés dans des structures syntaxiques typiques, on peut trouver de tels passages, en extraire des information sur l'organisation des termes entre eux[1], et les lier aux concepts sous la forme de *définitions encyclopédiques* ou tout au moins de gloses, comme illustré en figure 3.1. On espère ainsi donner une intelligibilité à l'ontologie, que cette intelligibilité autorise le contrôle des outils automatiques de conception des ontologies, ou facilite la compréhension du résultat final.

Mais on ne peut en toute généralité présupposer l'existence – ou l'accessibilité – de textes adaptés dans le domaine. Et, très souvent, les ontologies sont conçues *ad hoc*, et le lien au domaine textuel ne peut être obtenu qu'à la suite de la construction manuelle de définitions encyclopédiques. Mais dans ce cas, c'est la rationalité de l'approche qui est mise en cause : comment être sûr que les définitions ainsi construites forment un tout cohérent ? Une ontologie, comme un thesaurus, doit former un système interprétatif quasi-autonome. Il doit donc répondre à des normes strictes permettant à ses éléments de prescrire une interprétation qui, en plus d'être accessible, soit à coup sûr non ambiguë et non contradictoire.

Pour cela, comme dans les thesauri qui recourent prioritairement à l'organisation hiérarchique pour constituer des contextes locaux d'interprétation, les concepteurs d'ontologies misent sur la structuration offerte par la relation de spécialisation pour rendre le contenu de leur travail accessible à l'utilisateur final. Les ontologies sont donc le plus souvent présentées comme des hiérarchies de concepts et de relations, plutôt que comme des listes à plat, comme illustré dans la figure 3.2 qui reprend les vues offertes par différents éditeurs d'ontologies ou systèmes d'annotation. Et pourtant, de la même manière que cela peut ne pas réussir pour les thesauri, le recours à la subsomption ne garantit pas une interprétation correcte, même si cette subsomption est munie d'une interprétation formelle précise. En effet, on ne se place plus du tout du côté d'une interprétation naturelle, si ce n'est à laisser celui qui accède à l'ontologie interpréter ce qu'il voit d'une manière différente de l'interprétation formelle sous-jacente !

Dans certains cas, ceux où une application documentaire utilisant un thesaurus pré-existait au SBC ontologique, on peut vouloir importer la hiérarchie thésaurale. Ce calcul n'est pas mauvais au premier abord. On récupère ainsi un objet qui, conçu pour résoudre des problèmes de description documentaire dans le cadre d'une interprétation non formelle, bénéficie déjà d'une légitimité certaine et est d'un abord plus immédiat. Cependant, comme dans le cas des approches consistant à rattacher les notions ontologiques à des textes du domaine, on ne peut être sûr d'avoir les ressources adéquates à disposition.

De plus, et cela est particulièrement dangereux, le thesaurus est un artefact qui ne bénéficie absolument pas d'une précision équivalente à une ontologie. C'est d'ailleurs cela qui nous a déjà poussé à nous en défier. Les relations sémantiques entre les notions d'un thesaurus sont floues : la relation hiérarchique d'un thesaurus, en particulier, ne peut être interprétée comme dénotant l'inclusion entre des ensembles d'individus, puisqu'elle peut renvoyer aussi à la relation entre le tout et ses parties, entre un thème global et des sous-thèmes... Adapter directement une hiérarchie thésaurale risque donc très fortement d'aboutir à une utilisation abusive de la relation

[1]Pour des exemples concrets dans le domaine de structuration de terminologies, on peut se tourner vers [Le 00] et [MZB04].

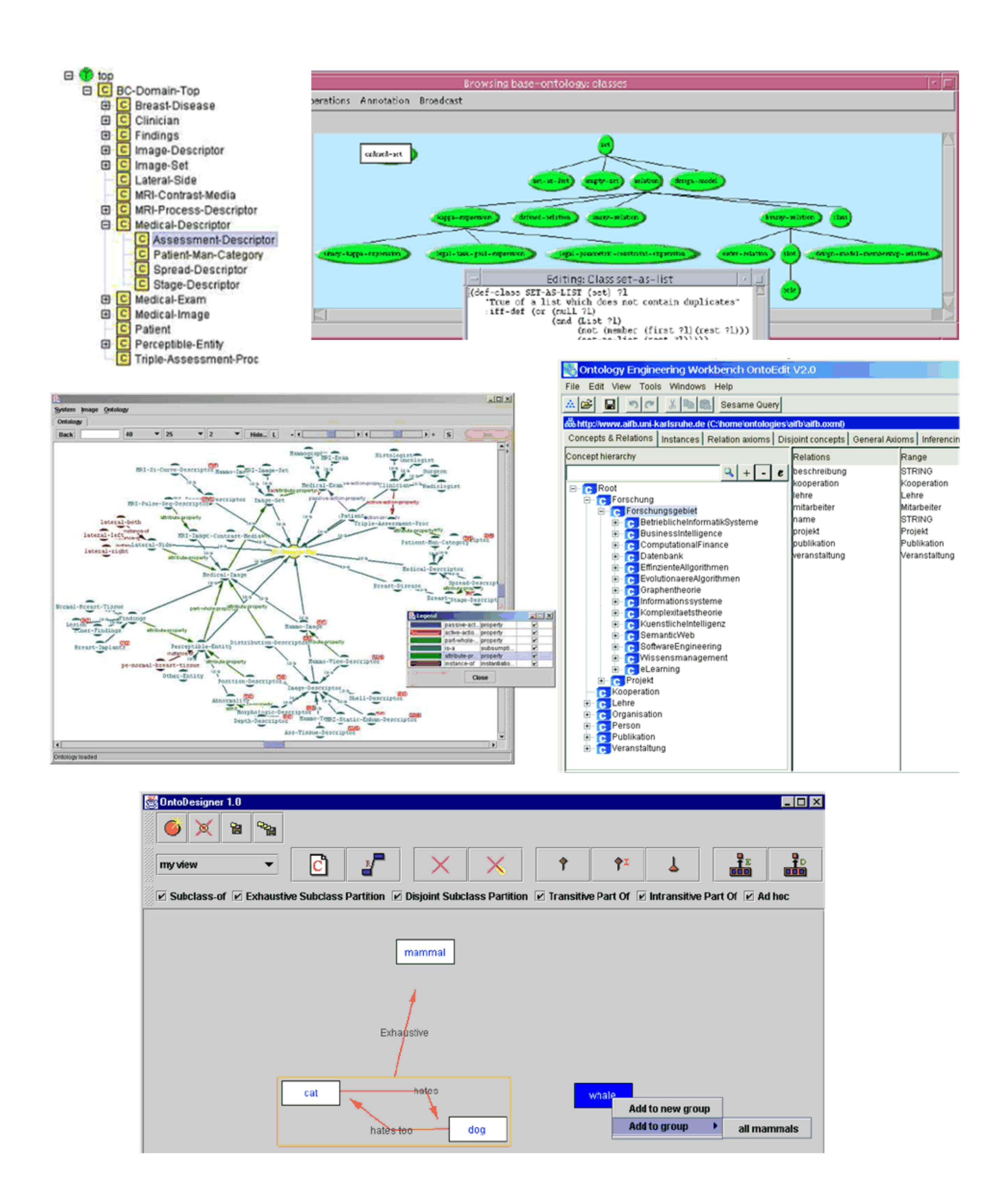


FIG. 3.2 Différentes visualisations hiérarchiques d'ontologies

OilEd[BHGS01], WebOnto[Dom98], Delogviz[HS03b], OntoEdit[SEA+02], WebODE[Arp00]

de spécialisation, ce que Guarino dans [Gua98b] appelle la *surcharge* de cette relation. On pourra par exemple trouver l'affirmation selon laquelle un objet physique *est* une quantité de matière, alors qu'en fait cet objet *est composé* de cette quantité.

Pour réutiliser un thesaurus dans un cadre d'indexation ontologique, on est donc obligé de faire un travail de re-formalisation et de « nettoyage », selon des critères comme par exemple ceux de Guarino que nous présenterons dans le chapitre 4. Les exemples de [AFS00, WSWS01, vMS+04] sont là pour montrer que cela est une opération difficile, qui demande beaucoup d'efforts. Et finalement, la hiérarchie obtenue peut être différente de celle du thesaurus initial.

Ces diverses approches ne garantissent pas à coup sûr l'accès à une signification non formelle adéquate pour les descripteurs employés dans le SBC. Les seules significations accessibles de façon certaine sont celles issues des spécifications formelles. On peut toujours essayer de les rendre plus accessibles, par l'emploi d'interfaces qui, dans la continuité de ce que l'utilisation des hiérarchies offre, « vulgarisent » les spécifications formelles. Poursuivant sur la lancée des langages de RC qui essaient d'utiliser des primitives épistémologiques relativement compréhensibles, présentées de manière plus ou moins informelles [GHB96, Kas02], différents outils de visualisation d'ontologies – la plupart couplés à des interfaces de conception de ces artefacts – proposent des moyens plus ou moins graphiques d'accéder à cette signification [CM01, LN03]. Certains outils proposent également de générer automatiquement des verbalisations des expressions définissant les concepts et les relations [RZS+99, KV03, KHGP04], comme en figure 3.3. On construit alors des expressions langagières à partir de termes associés aux primitives logiques et non logiques de la spécification de l'ontologie, et de règles syntaxiques qui reflètent celles qui s'appliquent au méta-langage. De telles solutions ont pu être appliquées à des règles d'inférences plus complexes que les expressions des logiques de description [HNM02], ou encore en association avec des moyens visuels classiques, comme dans le cas de l'extension de l'éditeur **Protégé** dédiée à la conception d'ontologies en OWL ([KFNM04], cf. figure 3.4 pour un exemple de spécification dans notre domaine audiovisuel) ou de l'éditeur d'ontologies développé dans le cadre du projet SHAKEN [TBC+02]. Mais quoi qu'il arrive, ces reformulations ne possèdent pas l'intelligibilité des expressions langagières « attestées » dans l'application. Et elles ne rendent toujours compte que des spécifications formelles, ce qui pose des difficultés, puisque celles-ci ne sont pas directement liées au domaine d'application, et donc peuvent être ressenties comme incomplètes par les utilisateurs.

Il faut donc aller plus loin, et munir les ontologies d'une interprétation naturelle qui recontextualise correctement concepts et relations dans les usages de l'application, même si ceux-ci ne sont pas directement accessibles lors de l'utilisation de l'ontologie.

Pour clarifier les rapports entre éléments langagiers et éléments conceptuels, Gilles Kassel dans [KP99, KAB+00] propose un modèle de représentation ontologique qui fait explicitement la distinction entre trois composants distincts d'un concept :

- les *termes* qui sont employés pour le désigner dans la langue ;
- la *notion*, son sens pour des humains immergés dans les pratiques de l'application ;
- les *objets* du monde qui forment la référence du concept.

La notion est évidemment ce qu'il convient d'expliciter correctement pour que le concept ait un sens accessible pour les utilisateurs de l'ontologie. Pour ce faire, il faut s'efforcer de doter les concepts de l'ontologie d'une définition en langue naturelle. De plus, pour structurer quelque peu cette signification, Gilles Kassel, s'inspirant d'une proposition que l'on retrouvera dans la section suivante de ce chapitre, propose d'introduire des *axes* permettant de regrouper parmi

Figure 2b: Detail of content from Figure 2a

RUBRIC	"botplastiek van ulna met bottransplantatie, niet gespecificeerd, inclusie:met aanbrengen van fixatiemateriaal"
PARAPHRASE	"plastic constructing of ulna by technique transplanting bone with immobilising using fixation device"
ENGLISH_RUBRIC	"boneplasty of ulna with transplant of unspecified bone, inclusion: with immobilising by means of fixation material"
SOURCE	"CVV" CODE "5-786.93b"
INTERMEDIATE REPRESENTATION	MAIN plastic constructing ACTS_ON ulna BY_TECHNIQUE transplanting ACTS_ON bone WITH immobilising BY_MEANS_OF fixation device
GRAIL	(SurgicalDeed which < isMainlyCharacterisedBy (performance whichG isEnactmentOf (SurgicalReshapingProcess whichG < LocativeAttribute Ulna hasSpecificSubprocess (SurgicalTransplantation which actsSpecificallyOn Bone)>)) isCharacterisedBy (performance whichG isEnactmentOf (SurgicalImmobilizing which hasSpecificPhysicalMeans FixationDevice))>) hasProjection (('CVV' schemeVersion 'default') code '5-786.93b' 'code')
GENERATED ENGLISH	Reshaping with transplanting of bone on ulna with immobilization using fixation device

Fig. 3.3 D'un langage de représentation formel à une représentation langagière (« Generated English ») dans le programme GALEN, extrait de [RZS+99]

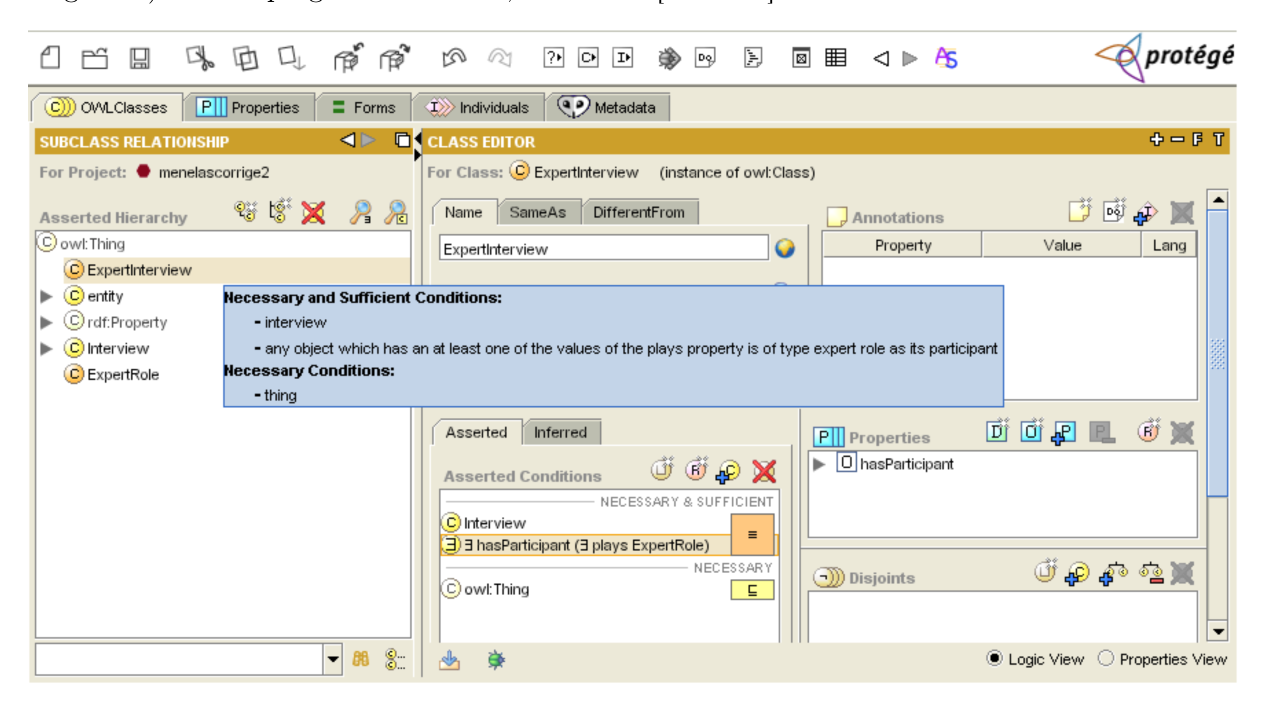


Fig. 3.4 Visualisation d'une classe définie dans **Protégé**

les spécialisations d'un concept donné ceux qui ont des critères de définition commun[2]. Ces regroupements permettent de mieux comprendre les raisons de l'introduction des concepts dans l'ontologie, ce qui est un atout en ce qui concerne leur utilisation future. Et, comme nous venons de le mentionner en page 84, cette spécification employant des éléments purement langagiers est ensuite doublée d'une spécification *semi-informelle* qui rend compte du sens formalisé des concepts et des relations en employant un langage semi-informel, dont la figure 3.5 donne un exemple.

Employee: [EP/SLD] an EMPLOYEE is a PERSON who WORKS ON BEHALF OF an EMPLOYER. [EP/RL] Every EMPLOYEE IS PAID BY the EMPLOYER who employs him. [SA] The concept EMPLOYEE is specialized in ENGINEER, RESEARCHER according to the nature of work realized by the EMPLOYEE.

Electronic document: [EP/SLD] An ELECTRONIC DOCUMENT is a DOCUMENT which HAS A SUPPORT electronic. [EP/RL] Every ELECTRONIC DOCUMENT HAS FOR FORMAT a FORMAT. [EP/LME] The ELECTRONIC DOCUMENTS are opposed to PAPER DOCUMENTS.

Works on behalf of; is employed by: [EP/SL] WORK ON BEHALF OF implies TO BE USED AS A RESOURCE BY. [DR & RR] An EMPLOYEE WORKS ON BEHALF OF an EMPLOYER.

FIG. 3.5 Définition de concepts et de relations selon la méthode OntoSpec

La compréhension de ces formulations reste évidemment délicate pour un non-expert, mais elle l'est moins que celle des axiomes formels qu'elles traduisent. Le problème des propositions de l'équipe du LARIA, de notre point de vue, réside plutôt dans la gestion des informations formulées en langues naturelle. Il n'y a en effet pas de guide très précis pour la création des axes et l'expression des différences : cela tient en grande partie au flou qui demeure quant à la manière dont on peut rendre compte de l'intension des concepts. Si on doit utiliser la langue, il serait bienvenu de restreindre la manière dont on l'emploie, de recourir à une théorie interprétative précise donnant aux définitions obtenues une signification prévisible. La possibilité de définir autant d'axes de spécialisation de concept que souhaité est également troublante : s'il existe plusieurs axes, peut-être pouvait-on introduire auparavant introduire un axe permettant de répertorier rationnellement ces différents points de vue de définition...

Les propositions de ces travaux de recherche vont néanmoins clairement dans le bon sens : si on suit cette approche, l'utilisateur peut disposer d'une explicitation lisible de la signification des entités de l'ontologie. Mais il semble que la clarté de l'engagement ontologique doit résulter d'une plus grande rigueur au niveau non formel, comme on va le voir à présent.

3.2.2 Des notions ontologiques normalisées

Dès qu'il s'agit de rendre accessible le sens des notions de l'ontologie, on a pu observer que l'utilisation de la langue naturelle était ce vers quoi l'on tendait : l'étiquetage des concepts par des signifiants langagiers est une pratique largement répandue, et l'on rencontre parfois des ontologies qui s'efforcent de verbaliser sommairement les expressions formelles complexes qui définissent leur notions, ou leur adjoignent des gloses textuelles classiques. Cette approche est rationnellement justifiée, puisque c'est généralement la langue qui est utilisée dans les domaines d'application dès qu'il s'agit d'expliquer des faits. Néanmoins, coupés de leur contexte, des éléments isolés comme les étiquettes des notions ontologiques n'ont pas de signification suffisamment déterminée : on est toujours à la merci des phénomènes de polysémie, par exemple. Ce n'est que lorsque le terme

[2]De fait, ces concepts sont en opposition, puisque les critères de définition permettent de les regrouper, mais, comme on va le voir dans ce qui suit, ce n'est que pour mieux les opposer entre eux.

est immergé dans son contexte d'application qu'il trouve une interprétation précise. Il manque donc un effort de rationalisation sémantique qui peut garantir que l'ontologie pourra prescrire des interprétations à la fois naturelles et sûres.

Pour répondre à ce besoin, Bruno Bachimont propose, dans une méthodologie introduite dans le cadre du projet MENELAS [ZC94], de contraindre l'utilisateur à un *engagement sémantique* munissant les primitives apportées par les ontologies d'une signification métier qui ait fait l'objet d'une *normalisation sémantique* [Bac00] . Comme nous le verrons dans le chapitre 4, cette proposition dépasse en fait le cadre de la simple interprétation de concepts et relations, puisqu'elle s'inscrit dans un processus complet de conception d'ontologies. Nous préférons néanmoins aborder dès ce moment les caractéristiques qui font d'elle un outil précieux pour l'accès à la signification de l'ontologie et des index qui sont conçus avec son aide.

Une sémantique différentielle pour les unités linguistiques

Tout part en fait des intitulés langagiers des concepts et des relations. Indépendamment des spécifications formelles de ces notions, il faut leur donner une signification informelle adéquate au domaine considéré. Pour cela, des textes sélectionnés pourraient constituer le contexte d'interprétation idéal. Néanmoins, leur rattachement à l'ontologie pose problème. Il faut donc se résoudre à recréer autour des notions un contexte artificiel, qui guide leur interprétation linguistique de manière satisfaisante. Quelle théorie sémantique peut-elle être utilisée à une telle fin ?

Bruno Bachimont se tourne vers la sémantique différentielle de François Rastier [RCA94]. Dans le paradigme différentiel, le sens est intra-linguistique : il se construit par des relations de similarité et d'opposition entre les unités du système linguistique. On ne se préoccupe pas du rapport des termes aux objets du monde, mais du rapport qu'ils entretiennent avec les autres termes.

En pratique, les unités du système linguistique trouvent une signification grâce à des traits sémantiques – les *sèmes* – qui vont permettre de définir des oppositions entre ces signifiants.

Les sèmes que l'on peut associer à une unité sont regroupables en deux catégories :

- les *sèmes génériques*, qui sont communs à d'autres unités ; ils permettent de regrouper les unités entre elles ;
- les *sèmes spécifiques*, qui permettent à une unité de se distinguer de celles avec qui on l'a regroupée grâce aux sèmes génériques.

Par exemple, si on se restreint au domaine cycliste[3], on va pouvoir associer, entre autres, le sème « réussit en montagne » au signifiant « grimpeur » et le sème « réussit en plaine » au signifiant « rouleur ». Par contre, elles vont partager les sèmes génériques « homme » et «dont le métier est de participer à des courses cyclistes», associés au signifiant « coureur cycliste ». Ces deux sèmes vont être génériques pour les unités « grimpeur » et « rouleur ». Par contre, les sèmes « réussit en montagne » et « réussit en plaine » sont des sèmes spécifiques qui permettent de distinguer ces deux unités l'une de l'autre.

Ces deux types de sèmes permettent de définir une unité par les identités et les différences qu'elle entretient avec les unités dont la signification est la plus proche. La généralisation de ces liens, permettant de relier toutes les notions entre elles, fait de cet ensemble un véritable système d'interprétation qui peut dès lors constituer un référentiel sémantique adéquat. C'est la position dans le système qui donne un sens à l'unité considérée.

[3]Les exemples proviennent d'une expérimentation conduite à l'INA en collaboration avec Raphaël Troncy et Estelle Le Roux, dans le cadre de leur thèses CIFRE respectives [Le 03, Tro04] – le lecteur peut se référer à la section 5.2.1 de ce manuscrit pour plus de détails. Nous avons préféré les conserver, en continuité avec les publications relatives à ce sujet que constituent [TI02, BIT02].

Il faut cependant signaler que ce sens différentiel est toujours dépendant d'un contexte, qui encadre l'interprétation, qui dans le cas de la sémantique différentielle est le processus d'attribution des sèmes aux unités linguistiques. La liste des sèmes associés à une unité est en effet variable : chaque occurrence de l'unité dans un contexte va enclencher un processus d'interprétation. Au cours de celui-ci, en fonction des sèmes inhérents de l'unité visée – ceux qui lui sont associés par défaut – et de ceux des unités qui se situent dans son contexte, on opère une (ré-)attribution des traits sémantiques : certains sèmes d'une unité seront activés, alors que d'autres seront désactivés.

Toute la difficulté vient alors du fait que pour obtenir une primitive réellement exploitable il faut supprimer cette dépendance au contexte : on doit opérer une *normalisation sémantique*. Cela se fait en ramenant la dimension du contexte au domaine considéré, action qui autorise une assignation plus stable de sèmes aux unités langagières. L'interprétation est de fait encadrée, ce qui permet de définir la signification recherchée. D'une simple description linguistique, on passe à un sens véritablement prescriptif, qui permet d'envisager le rapport d'une primitive, l'unité de sens que nous recherchons pour une description, à ce qu'elle est supposée représenter dans le domaine.

L'ontologie différentielle

Pour normaliser le sens variable des unités étudiées, il faut donc énoncer des consignes pour l'attribution des sèmes génériques et spécifiques, consignes qui sont nécessairement liées à la pratique visée par l'ontologie. Il faut que ceux qui accèdent à l'ontologie assignent un sens aux termes qui soit bien celui auquel aboutirait le processus d'interprétation. Cela se fait en fixant et en structurant le réseau d'identités et de différences que l'on avait évoqué dans le paragraphe précédent.

Tout d'abord, on note qu'une notion se définit d'abord par un lien de subsomption avec une autre notion : celle dont elle hérite des sèmes génériques. Par exemple, on pourra affirmer que `coureur cycliste` subsume `grimpeur` et `rouleur`. En pratique, chaque notion doit être caractérisée par ses similarités et ses différences avec les notions qui sont placées dans son voisinage proche : la notion-parente[4] et les notions-sœurs.

Pour expliciter ces information, [Bac00] propose quatre principes, inspirés de la méthode aristotélicienne de définition des essences par genre proche (*genus*) et différence spécifique (*differentia*) :

- le principe de *communauté avec le père* (dans nos travaux, on retrouvera les dénominations *« similarity with parent »* ou **SWP**) : on explique pourquoi le fils hérite des propriétés de la notion qui le subsume ;
- le principe de *communauté avec les frères* (*« similarity with siblings »* ou **SWS**) : on donne un axe sémantique, une propriété admettant plusieurs valeurs exclusives qui permet de comparer les notions d'une même fratrie ;
- le principe de *différence avec les frères* (*« difference with siblings »* ou **DWS**) : on indique ici la caractéristique qui permet de distinguer la notion considérée de ses notions-sœurs ;
- le principe de *différence avec le père* (*« difference with parent »* ou **DWP**) : on explicite finalement la différence qui permet de distinguer le fils du père, différence résultant généralement des traits ayant permis de caractériser distinctement la notion du reste de sa fratrie.

[4] Bruno Bachimont démontre que cette hiérarchie a une structure d'arbre : l'héritage multiple est interdit, dans la mesure où l'on associe nécessairement aux notions d'une fratrie – et par conséquent aux notions qui les spécialisent – des sèmes spécifiques qui sont en opposition.

Par exemple, si l'on veut fixer ces principes pour les notions qui spécialisent directement la notion `Personne` dans l'ontologie du cyclisme, on obtient le tableau 3.2.

`Personnel Épreuve`	
SWP :	c'est un individu humain
SWS :	une propriété précise la raison de la personne sur les lieux de l'épreuve
DWS :	est accrédité par la direction de l'épreuve
DWP :	joue un rôle particulier par rapport à l'épreuve : la personne est accréditée par la direction de celle-ci
`Personnel Équipe`	
SWP :	c'est un individu humain
SWS :	une propriété précise la raison de la personne sur les lieux de l'épreuve
DWS :	est employé par une des équipes participantes
DWP :	joue un rôle particulier par rapport à l'épreuve : la personne est employée par une équipe cycliste participante à l'épreuve
`Spectateur`	
SWP :	c'est un individu humain
SWS :	une propriété précise la raison de la personne sur les lieux de l'épreuve
DWS :	n'est ni accrédité par la direction de l'épreuve, ni employé par une des équipes participantes
DWP :	joue un rôle particulier par rapport à l'épreuve : se contente d'y assister

TAB. 3.2 Principes différentiels associés aux spécialisations directes de *Personne*, extrait de [TI02]

Au final, les similarités et les différences collectées sur le chemin qui mène d'une notion à une autre permettent d'inter-définir celles-ci sans qu'il y ait besoin de se tourner vers un contexte donné. En particulier, si on part de la notion la plus générique de la taxinomie ainsi construite, celle qui ne possède pas d'autre propriété que de relever du contexte applicatif, on peut construire la signification « absolue » (dans ce contexte applicatif) de chacune des unités langagières employées. On est passé d'un signifiant dont la signification n'était pas assez contrainte à un véritable concept, dont la signification, ancrée dans le domaine, est invariable, et peut donc fonctionner comme une primitive exprimant une connaissance.

On a donc ainsi créé une hiérarchie de notions interprétables de manière naturelle par les utilisateurs humains du SBC. Ces utilisateurs pourront donc mieux appréhender la signification des descripteurs qu'ils emploient pour créer les index. Il suffit de considérer que ce réseau de spécification sémantique est une partie intégrante de l'ontologie : l'*ontologie différentielle*.

Contexte d'interprétation et indexation

Nous avons mis en pratique les propositions théoriques de Bruno Bachimont tout au long de notre travail à l'INA, en particulier dans le cadre du projet OPALES. Raphaël Troncy et nous-mêmes avons développé un outil de création et de visualisation d'ontologies, **DOE** – pour *Differential Ontology Editor* – qui répond aux spécifications énoncées dans [Isa01]. Cet outil, dont nous reparlerons en détail dans le chapitre 4, a été inséré dans la plate-forme d'annotation d'OPALES [ICG+04], de sorte que l'utilisateur, lorqu'il crée un index conceptuel, puisse avoir accès à l'interprétation métier de la notion qu'il souhaite employer, et, le cas échéant, en sélectionne une plus appropriée à ses intentions. La figure 3.6 montre l'utilisation de l'ontologie différentielle lors de l'indexation, dans le contexte de l'application d'OPALES dédiée à la petite enfance.

Grâce à de telles propositions méthodologiques, on est à même de faciliter le travail de l'indexeur, et, plus généralement, la tâche de tous ceux qui ont accès aux concepts et aux relations

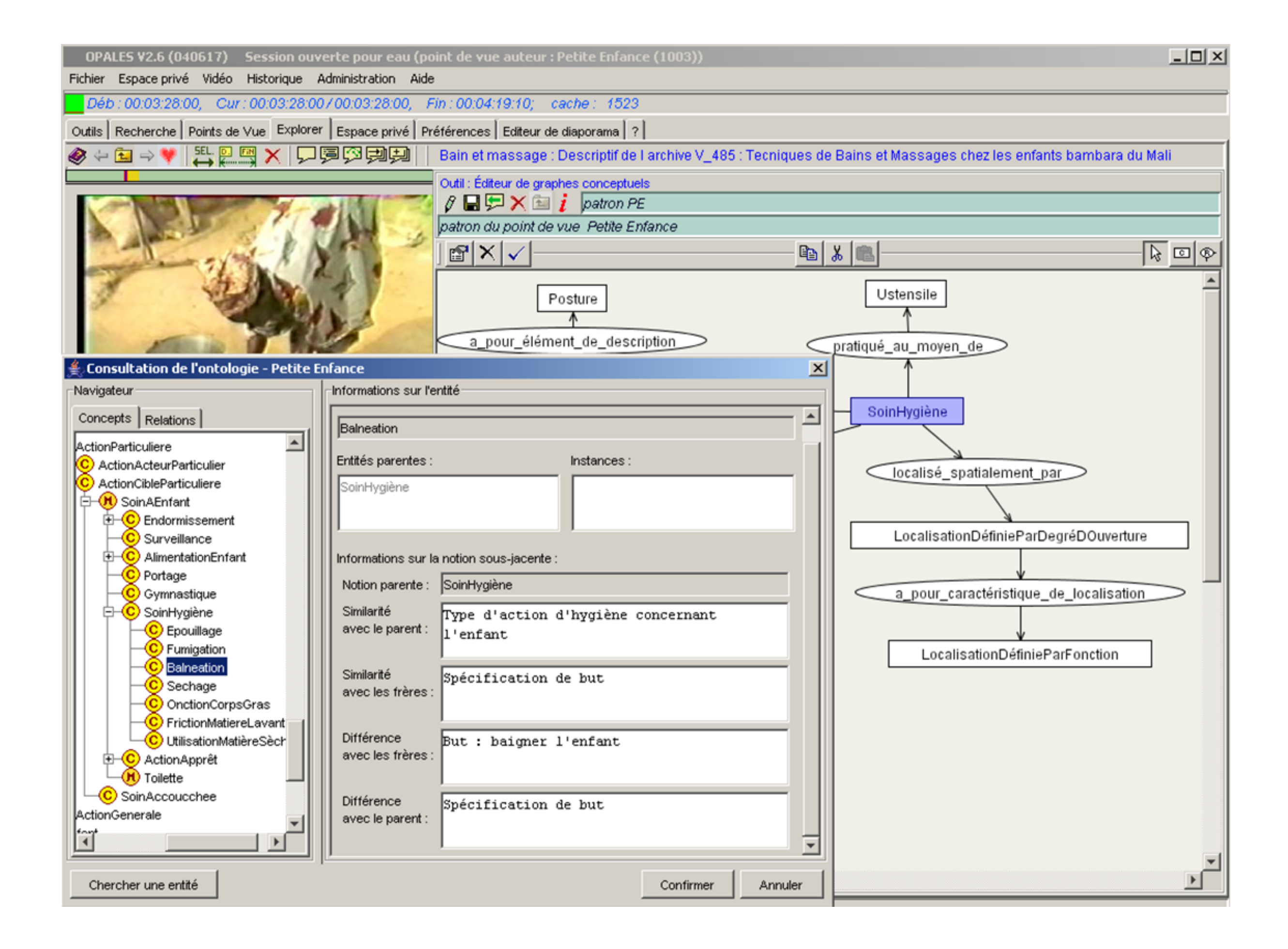


FIG. 3.6 Utilisation de l'éditeur d'ontologies pour la création d'un GC index dans OPALES

introduits par l'ontologie. Celle-ci *fait* désormais *sens*, dans toutes les acceptions de cette expression.

3.3 Assister la formulation des descriptions

Et pourtant, accéder à une signification intelligible des éléments de la description ne suffit pas à garantir une indexation qui réponde à tous nos critères de pertinence. Nous avons en effet, pour assurer la continuité sémantique, besoin d'une certaine cohérence des index, à la fois entre eux et avec les besoins applicatifs. Or, si elle donne un contexte d'interprétation satisfaisant pour les notions de l'ontologie, l'ontologie différentielle ne suffit pas à recréer un contexte d'emploi suffisamment précis. En d'autres termes, elle permet de contrôler l'interprétation des index, mais elle ne prescrit pas d'utilisation particulière en matière de conception de connaissances structurées. Cela peut se révéler gênant pour le créateur des index, d'autant plus qu'ici la reformulation des contenus doit aboutir à un résultat plus complexe que celui d'une indexation classique.

Indirectement, pourtant, les ontologies ont une incidence sur la forme des index réalisés, *via* les spécifications formelles qui permettent de contrôler la structure des descriptions. Par exemple, on a vu que les domaines formels des relations permettent de contrôler les individus qu'elles relient. En conjugaison avec d'autres connaissances formelles comme la hiérarchie de subsomption ou des axiomes comme l'exclusion entre concepts, on peut donc contrôler les assertions relationnelles. Le problème est que ce contrôle reste local. En principe, certains langages de RC proposent des primitives permettant d'interdire certaines configurations plus complexes. C'est le cas des logiques de descriptions qui autorisent l'élaboration de descriptions comportant des négations de formules complexes grâce au constructeur `not` de complémentation de concept. On pourrait ainsi définir la relation `aPourTraitSaillant` comme ayant pour domaine l'ensemble des objets audiovisuels à l'exception des ceux qui font l'objet d'une décomposition dans la BC : on obligerait ainsi l'indexeur à spécifier quels sont les éléments de cette décompostion auxquels s'applique cette propriété.

$$\texttt{aPourTraitSaillant} \sqsubseteq \texttt{domain}\ (\texttt{and AVObjet}\ (\texttt{not}\ (\texttt{some contient.AVObject})))$$

Mais ces possibilités restent peu utilisées. En grande partie, cela tient à la complexité induite par de tels mécanismes, que ce soit pour les systèmes de raisonnement ou bien pour les humains qui doivent concevoir les spécifications formelles. Mais on peut aussi considérer que l'on n'a toujours pas là de réponse à la question de la complexité de la création des index. Bien au contraire : on rajoute des spécifications que l'indexeur doit assimiler, tout en ne donnant pas de consigne « positive » sur la structure générale de ce qu'il doit effectivement formuler.

Le risque est donc grand de voir la variabilité intra (ou inter-) indexeur atteindre un niveau élevé : sans consignes, on peut retomber dans la situation observée pour les systèmes documentaires utilisant la langue non contrôlée. Dans l'absolu, les SBC permettent d'employer des langages riches et de produire une multitude de combinaisons. Ceci peut évidemment nuire à la continuité sémantique et à la bonne utilisation du système d'indexation et de recherche. Comment peut-on être certain que les index répondront aux besoins de la recherche, et en particulier que les variations observées ne perturberont pas les résultats des processus de recherche, dont nous avons vu qu'en théorie ils pouvaient grandement bénéficier du recours à des ressources ontologiques riches ?

Il faut, là encore, que le recours à un cadre complexe d'indexation ne coupe pas l'utilisateur du contexte traditionnel que constituent les pratiques de l'application visée par le système. Nous allons voir comment peuvent être mises en places des solutions à même d'assister le travail des indexeurs.

3.3.1 Prescrire le contenu des index

Ce dont il s'agit ici, et dans le cadre de la continuité sémantique, c'est, après avoir prescrit une interprétation des index en donnant leur vocabulaire muni d'une signification claire, de leur prescrire un contenu, prescription qui concerne aussi bien leur substance que leur forme. Il faut déterminer une structure informationnelle pertinente pour l'application, et voir comment on peut l'exploiter pour faciliter le travail des différents utilisateurs. Il faut noter que dans cette thèse qui vise plus à la représentation et la gestion des connaissances dans un cadre ontologique, nous n'allons pas nous intéresser à la manière d'obtenir de telles données elles-mêmes, ce qui relève d'un travail d'acquisition de connaissances pure. Nous allons plutôt exhiber la forme que ces données peuvent prendre, et, surtout, quel est l'usage qu'on en fait dans un SBC.

L'assistance à la création des index dans les solutions documentaires classiques

Le problème du contrôle éditorial des indexations a évidemment fait l'objet de considérations avant l'introduction des systèmes à base de connaissances. Le problème de la variation des index ne s'appliquant pas qu'au vocabulaire employé, mais aussi à l'organisation des informations, il s'agissait de faire quelque chose pour que les systèmes documentaires limitent aussi les risques d'incohérence à ce niveau. Cela est particulièrement important dans les systèmes documentaires autour desquels travaille un grand nombre de documentalistes, et pendant une assez longue durée, comme c'est le cas à l'INA. Dans de telles situations, les risques de variation dans le processus et les résultats de l'indexation sont élevés.

Tout d'abord, il faut observer que les langages documentaires traditionnels n'ont pas vocation à proposer des expressions privilégiées, pas plus qu'ils ne peuvent, en l'absence de spécifications formelles riches, contraindre l'utilisation du vocabulaire au sein des expressions construites. Il faut donc se tourner vers l'environnement du langage documentaire, que ce soit le système lui-même ou son environnement organisationnel plus large.

De fait, on fait souvent en la matière confiance au savoir-faire de celui qui indexe. C'est son expertise dans le domaine, ainsi que sa connaissance du système documentaire, sur le plan technique et méthodologique, qui sont garantes de la formulation correcte d'une « bonne » interprétation. Ce pourquoi on considère qu'il vaut mieux que les index soient réalisés par des sujets ayant bénéficié d'une formation approfondie, ou qui ont à disposition les ressources qui leur permettent de faire face aux interrogations qui ne manqueront pas d'apparaître pendant l'indexation. Il faut noter que les ressources dont il est question sont aussi bien d'origine humaine – connaissances des autres indexeurs, expérience acquise par le sujet indexant – que matérielle. Par exemple, l'INA, comme beaucoup d'institutions s'efforçant de capitaliser leurs connaissances, a conçu des guides méthodologiques recensant les bonnes pratiques susceptibles de répondre à ses besoins. Ces guides sont présents sous la forme de « bibles » d'indexation [Pic96], ou d'indications réparties entre les différents descripteurs dont l'emploi peut poser problème. Des notes d'application extérieures au thesaurus, et plus complètes que celles-ci, en quelque sorte... L'INA s'intéressant à la caractérisation précise des programmes de télévision, on trouvera toute une série de recommandations concernant l'emploi de plusieurs catégories d'objets audiovisuels pour décrire un programme donné. Ainsi, à la rubrique « mini-programme », en plus des précisions sur la signification du descripteur correspondant, on trouve des règles d'usages qui indiquent quelles sont les catégories plus techniques qui lui sont typiquement associées : interview, animation, montage d'archives et autres.

Il existe cependant des solutions plus formalisées, des *grilles d'indexation* qui indiquent directement, pour différents thèmes ou événements, les descripteurs qui doivent apparaître dans

l'index. Par exemple, on peut trouver une grille relative à un événement de type « accident médical », illustrée dans le tableau 3.3. Ces structures recommandent des descripteurs issus du thesaurus de l'INA, aussi bien que du texte libre quand il n'existe pas de descripteur adapté. Elles permettent à l'indexeur de savoir quelle est l'information qui est pertinente, ce qu'il doit éventuellement modifier, retrancher ou ajouter selon le contenu qu'il rencontre. En fait ces grilles doivent être comprises comme des candidats-index, à adapter au cas particulier que constitue chaque document indexé. Ainsi, dans notre exemple, l'indexeur doit pouvoir remplacer le descripteur `centre hospitalier` par `clinique` si le besoin s'en fait sentir.

ACCIDENT MEDICAL		
DEL	Pays si différent de France	
	Ville	
DET	Médecine	
	Centre hospitalier	Nom de l'hôpital en champ Texte
	Accident chirurgical	
	Erreur médicale	

TAB. 3.3 La grille d'indexation d'un accident médical

Nous devons noter que si les grilles ressemblent, par leur aspect, à un formulaire, il y a pourtant des différences subtiles. Tout d'abord, elles ne bénéficient pas d'une structure relationnelle plus précise que celle offerte par les champs du « formulaire » d'indexation introduit pour l'indexation thématisée à l'INA (cf. p.29). En ce sens, elles ne proposent pas de schéma véritablement adapté à chaque besoin applicatif : on ne sait pas si le centre hospitalier est considéré comme le responsable de l'accident, ou comme un simple lieu, alors que cette information peut sembler importante dans ce cas. De plus, les grilles proposent déjà certaines des informations que l'on retrouvera dans l'index, alors qu'un formulaire se contente de donner une structure, sans mentionner les valeurs thématiques qui s'inséreront dans cette structure[5]. Une grille se place résolument dans une approche prescriptive : un formulaire peut par exemple contraindre certains champs – ce qui n'est pas le cas de la structure de formulaire de l'INA – mais il ne propose pas de valeur par défaut.

Il est acquis que ce sont de telles solutions, inscrites matériellement à l'aide du langage d'indexation et de lui seul, qui doivent nous inspirer, puisque nous nous sommes restreint dans le chapitre 1 à n'introduire que des solutions qui concernent le *support* de l'indexation. Il n'est pas question de proposer ici des réflexions sur la manière de former des documentalistes à l'indexation à base de connaissances, même dans le cadre d'une application précise ...

En somme, une grille constitue une référence pour les index réellement produits, en donnant un exemple de l'information qui a été jugée pertinente pour les recherches à venir. On a là un moyen d'agir sur la cohérence et la pertinence des index à un niveau différent de celui de la prescription d'une interprétation pour les descripteurs, puisque on fait plus que donner une signification. Cet objet est par ailleurs plus intéressant pour l'utilisateur voulant créer des index qu'un simple mécanisme de contrôle du contenu et de structure. On peut d'ailleurs remarquer que ces trois solutions – prescription d'interprétation, prescription d'indexation, contrôle des index produits – ne sont pas redondantes. L'emploi de préconisations de contenu doit bien évidem-

[5]Ou s'il le fait, ça n'est qu'indirectement et de manière plutôt figée, par l'intermédiaire de l'énumération d'une liste des valeurs possibles.

ment s'appuyer sur des significations claires données aux descripteurs qu'elles mobilisent. Et les mécanismes de contrôle sont nécessaires pour valider les éventuelles modifications qui seraient apportées à l'index préconisé. Et, de fait, on observe dans les systèmes à base de connaissances publiés des approches qui empruntent à ce qui se fait dans le domaine documentaire et cherchent à conjuguer ces différents aspects.

Des démarches d'assistance à la création des index dans un cadre ontologique

De manière générale, le problème de la création des assertions dans les SBC utilisant des référentiels ontologiques n'a pas fait l'objet de beaucoup de recherches ; en tout cas, rares sont ceux qui en font explicitement mention et qui proposent des mécanismes à même de compenser la complexité inhérente à ces systèmes. L'accent est le plus souvent mis sur les interfaces de visualisation des assertions et des documents qui y sont liés[6] que sur la mise en place de véritables méthodes d'assistance à l'utilisateur[7].

Parmi les approches proposées pour guider le processus d'indexation, le formulaire[8] est celle qui apparaît le plus souvent, sous des formes et avec des emplois divers. On retrouve donc des solutions proches de ce qui a été proposé dans le cadre des systèmes classiques dès qu'il s'agit de gérer des index dont la structure est complexe, solutions que nous avions déjà évoquées au chapitre 1.

Ainsi, dans le SBC CONCERTO de [ZBB+99, Zar00], le « répertoire des éléments ontologiques » contient, outre les concepts employés pour créer les annotations, des *templates* d'inspiration linguistique. Ces schémas de description sont associés à divers « événements », qu'ils décrivent à l'aide d'attributs spécifiques, proches des *cas* sémantiques [Fil68]. On aura par exemple un template permettant de caractériser une situation « être présent quelque part », par la mention d'un sujet – obligatoire, d'un objet – facultatif . . . Celui-ci sera spécialisé pour décrire une situation plus précise, « être présent avec quelqu'un », dont les attributs et leurs modalités pourront être différents, et qui pourra à nouveau être dérivé en un template spécifique à une application donnée. Chaque template fournit la liste des éléments qui sont nécessaires à sa caractérisation complète, comme montré en figure 3.7.

On retrouve là une vision proche de celle des *frames* [Min81], où un concept est introduit comme une construction composée d'attributs qui prennent des valeurs spécifiques à chaque instance. Le problème est que cette vision est plutôt restreinte quant à ses possibilités du point de vue des relations entre éléments : on ne peut exploiter des structures complexes de façon flexible, puisque les liens tranversaux entre entités du domaines ou entre entités et événements sont en retrait par rapport aux objets fondamentaux de la modélisation que sont les événements. On sera donc moins tenté par remettre en cause la structure préconisée, puisqu'elle est attachée

[6] [QKH03] et [Svd+04] donnent un aperçu des réflexions sur ce domaine, dans le cadre du web sémantique.

[7] Nous excluons tous les travaux qui visent à la population automatique de base de connaissances [Le 03, SCB+04, HS03a]. S'ils parviennent à masquer la complexité des SBC en soulageant l'utilisateur de la tâche de création des connaissances, ils se placent dans un cadre résolument différent du nôtre. On pourra d'ailleurs remarquer qu'à peu près tous utilisent pour guider les algorithmes d'extractions de connaissances des structures pressenties comme étant celles qui devront apparaître dans la BC. Et que dans ce cas, c'est en quelque sorte le système lui-même qui a besoin d'une assistance comparable à celle que demanderait un utilisateur humain . . .

[8] L'usage en anglais est de recourir au terme *template*, qui peut se traduire par celui de *patron*. Néanmoins, et pour marquer la différence avec les *patrons d'indexation* que nous proposons par la suite, nous présenterons comme des formulaires ce qui a pu être introduit comme *template*. En effet, les structures désignées par ce terme sont plus proches d'une structure linéaire à remplir que d'une structure articulée pouvant être directement imitée pour la création des index, ce que sont nos *patrons*.

```
name: Exist:BePresent
father: Exist:
position in H_TEMP: 2.3 ; NL description: 'Be Present Somewhere'
      EXIST         SUBJ        var1: (var2)
                    [OBJ        var3: (var4)]
                    [SOURCE     var5: [(var6)]]
                    !(BENF)
                    [MODAL      var7]
                    [TOPIC      var8]
                    [CONTEXT    var9]
                    { [ modulators ], ≠abs }

      var1  =  var3  =  var5  =  <human_being_or_social_body>
      var7  =  <action_name>
      var8  ≠  <property_>
      var9  =  <event_> | <action_name>
      var2, var4, var6  =  <physical_location>
name: BePresentWithSomeone
father: Exist:BePresent
position: 2.32  ; NL description: 'Be Present Somewhere with Someone'

      EXIST         SUBJ        var1: (var2)
                    OBJ         var3: (var4)
                    [SOURCE     var5: [(var6)]]
                    !(BENF)
                    [MODAL      var7]
                    [TOPIC      var8]
                    [CONTEXT    var9]
                    { [ modulators ], ≠abs }

      var1  =  var3  = var5  =  <human_being_or_social_body>
      var7  =      <action_name>
      var8  ≠      <property_>
      var9  =      <event_> | <action_name>
      var2, var4, var6  =  <physical_location>
      var2  =  var4
```

FIG. 3.7 Le template « Be present » et une de ses spécialisations, extrait de [Zar00]

de manière très forte – il s'agit presque d'une définition, qui serait donnée là implicitement – à la notion qu'elle représente. De plus, si on a une prescription syntaxique claire, on n'a pas véritablement d'indication complète pour le contenu : les attributs des événements présentés dans les formulaires ont des valeurs contraintes, mais ces contraintes – et c'est normal – font référence aux concepts les plus généraux du répertoire ontologique autorisés pour les champs concernés. On a donc une structure de description qui se comporte plus comme un récapitulatif des contraintes ontologiques que comme une prescription d'un contenu pertinent à un niveau immédiatement compréhensible par l'utilisateur. Celui-ci se retrouve en effet confronté dans sa tâche à l'ensemble de la hiérarchie des concepts, alors que, comme on le verra plus tard, la donnée d'un concept typique de l'application restreint la recherche au voisinage de celui-ci.

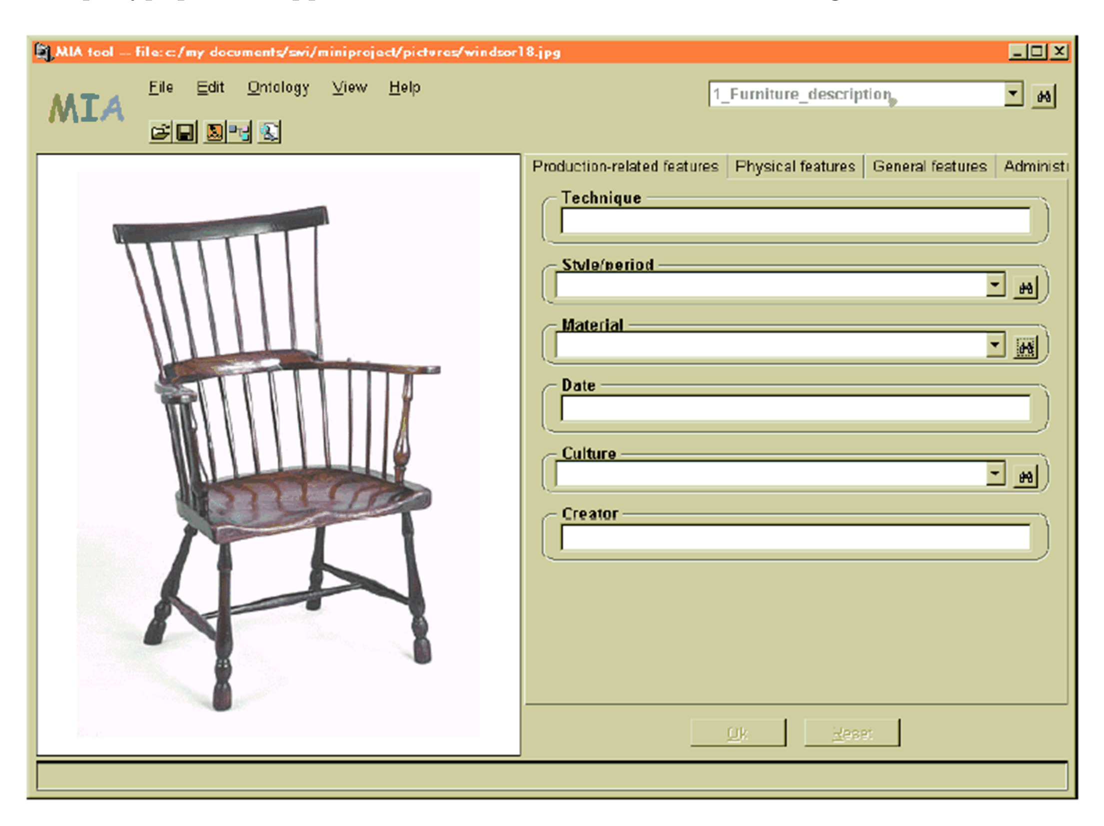


Fig. 3.8 La saisie de méta-données dans le projet **MIA**, extrait de [SBC$^+$02]

Plus proches des possibilités expressives et inférentielles offertes par les ontologies sont les propositions de l'équipe de l'université d'Amsterdam [SBC$^+$02, SDWW01, HSWW03], que nous avons déjà évoquées. Dans les plate-formes d'annotation de documents visuels que ces travaux présentent, il est tenu compte du contrôle éditorial qui a cours dans les pratiques documentaires visées par ces applications. Les concepts de l'ontologie, comme dans [ZBB$^+$99], sont à rapprocher de frames, mais cette fois-ci les attributs sont eux aussi des primitives de connaissances relevant de l'ontologie utilisée par le système, et non d'un schéma général de description. De fait, on a affaire ici à des adaptations des approches thésaurales classiques dans le cadre ontologique : [SBC$^+$02] introduit pour la saisie des méta-données documentaires sous forme de connaissances un *template* de description pour les objets à décrire, qui ressemble fort aux notices documentaires que l'on peut trouver dans les systèmes traditionnels (cf. figure 3.8). On crée ainsi une interface d'annotation qui concrétise en termes de concepts et de relations de l'ontologie une manière

d'organiser les données à saisir, et qui peut assister la tâche des utilisateurs. Cette démarche a été souvent reprise, dans [MSD00] ou [DDH$^+$04] par exemple. L'utilisation d'un tel schéma de description se révèle consensuelle et pertinente, puisque les experts du domaine qui ont été confrontés au prototype présenté dans [SBC$^+$02] s'y sont adaptés sans difficulté notable.

Cependant de telles solutions manquent de souplesse, en particulier au niveau de l'expressivité relationnelle. En effet, si les champs du formulaire sont modifiables, on se rend compte que cela ne concerne que les valeurs qu'ils peuvent prendre. On est ainsi libre de modifier les concepts employés dans une description, mais pas les relations qui les unissent. On ne peut donc pas véritablement changer la structure du formulaire, si ce n'est à en créer un autre. Et si le schéma de description jugé pertinent pour l'application admet des variations importantes dans quelques cas particuliers, on ne peut pas vraiment en tenir compte ; l'outil d'assistance éditoriale se mue alors en une contrainte qui peut rebuter l'utilisateur.

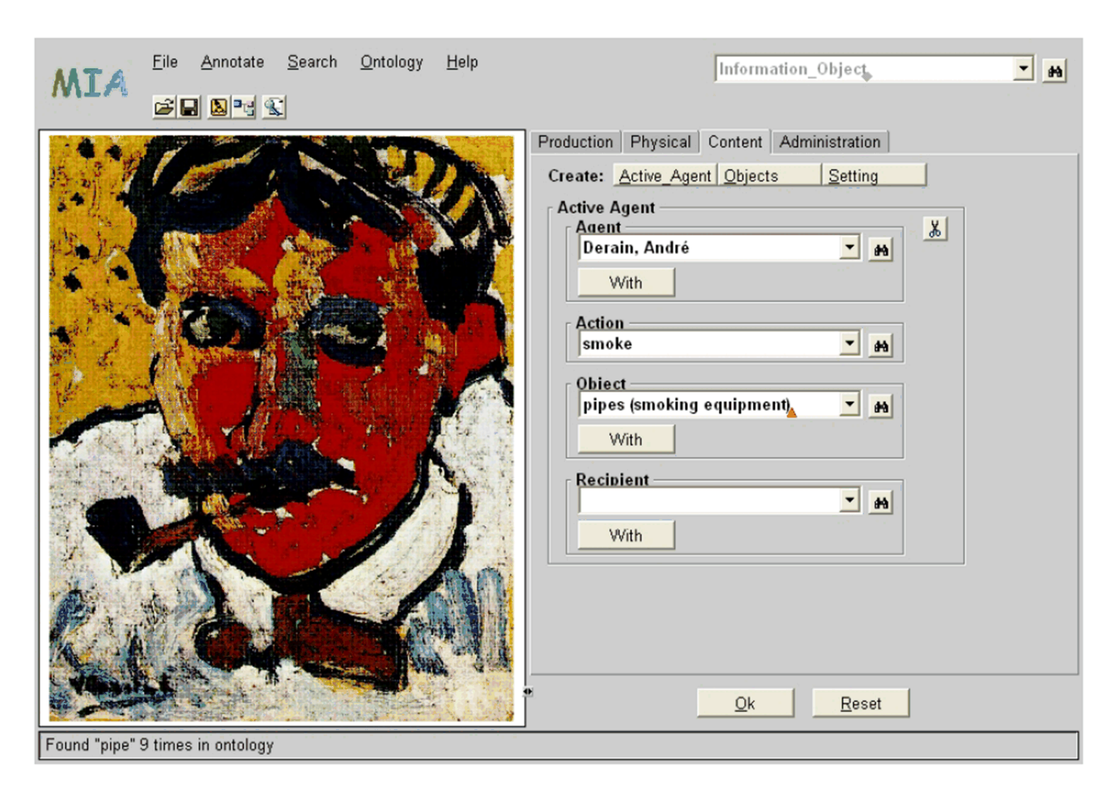


FIG. 3.9 La description du contenu de l'image dans le projet **MIA**, extrait de [HSWW03]

De fait, Schreiber et ses collègues ont tenu compte de cette difficulté, et proposé, pour la description du contenu des images proprement dit, des formulaires beaucoup plus souples. Comme la figure 3.9 le montre, leurs champs peuvent en effet contenir des structures de connaissances complexes qui sont saisies à leur tour dans des formulaires. Si l'utilisateur veut ajouter un élément de description, il n'aura qu'à associer au champ de formulaire correspondant un concept de l'ontologie, et l'interface présentera alors de nouveaux champs correspondant aux attributs du concept sélectionné. Cependant, on reste encore dans un cadre relativement figé : si le schéma de description jugé pertinent pour l'application admet des variations importantes dans quelques cas particuliers, on ne peut pas vraiment en tenir compte, et l'outil d'assistance éditoriale se mue en une contrainte qui peut rebuter l'utilisateur. Les auteurs reconnaissent par exemple que si leur schéma est adapté à la description d'*actions* comprenant un seul *agent*, il faudra recréer

une interface si une action en comportant plusieurs apparaît.

De façon générale, un tel constat est évidemment valable pour la description du contenu des documents audiovisuels, une tâche souvent plus complexe que la caractérisation de leur contexte de production, qui utilise habituellement des ensembles de méta-données relativement stables et adaptés à la saisie par formulaires. Le contenu audiovisuel, on l'a vu, est en particulier propice à une profusion de liens entre ses éléments, qui prohibe toute approche trop rigide de ce point de vue.

Si on peut faciliter la tâche des utilisateurs en ordonnant les propriétés des objets qu'ils sont censés décrire, cette assistance ne doit pas nuire à l'obtention des fonctionnalités – en particulier **F1** – nécessaires à une indexation correcte. Comme évoqué dans [BLGP03], on a souvent le réflexe de simplifier abusivement l'indexation pour diminuer les problèmes de variabilité des descriptions et améliorer les performances des systèmes d'information. Cela peut être dangereux : les seules contraintes devant effectivement s'appliquer aux index construits sont celles liées aux concepts et relations de l'ontologie, et non celles dérivant de l'emploi d'interfaces données. Une fois assuré le contrôle des index par le SBC au niveau même de la connaissance, il n'est en principe pas utile de rajouter une interface de contrôle éditorial strict. Une préconisation souple de la structure des descriptions semble beaucoup plus indiquée, surtout dans le cas de l'indexation des documents audiovisuels.

3.3.2 Patrons d'indexation

Des structures pour initier l'indexation

Afin d'assister l'indexation sémantique, il faut donc proposer une structure qui ait une valeur prescriptive plutôt que contraignante. Il faut également que cette structure mette l'accent sur l'emploi des relations de l'ontologie. Dans le cadre du projet OPALES, ces besoins se faisaient sentir de façon importante, puisque les communautés en présence souhaitaient décrire des faits dont le niveau de structuration est élevé : événements impliquant plusieurs participants faisant possiblement l'objet d'une description, rapports complexes entre des documents structurés, les éléments de cette structure et le contenu thématique associé à ces éléments.

Pour remédier à ce problème, la façon la plus naturelle est de fournir une première forme, approximative, d'index sémantique. L'utilisateur se voit ainsi proposer une information dont il pourra directement s'inspirer pour effectuer son travail. En fait, sa tâche est d'autant plus simplifiée que cette forme est intégrée au processus d'indexation, en en constituant par exemple le point de départ effectif. Pour OPALES, de tels index génériques ont été créés, et employés lors de l'indexation sous l'appellation de *graphes patrons*[9]. Pour chaque point de vue, les concepts et relations de l'ontologie correspondante sont mobilisés dans un graphe conceptuel censé refléter un index typique. Cette approche peut sembler proche de celle des formulaires, mais elle marque une progression sur deux points essentiels : on utilise les relations du domaine applicatif, et on prescrit un contenu conceptuel spécialisé.

Par exemple, pour le point de vue relatif à la petite enfance, les actions décrites
- impliquaient souvent une *mère* – souvent caractérisée par sa *posture* corporelle – et son *enfant*,
- étaient pratiquées à l'aide d'*ustensiles* et

[9]Pour se démarquer du choix particulier des GC pour la représentation des connaissances dans OPALES, nous utiliserons ensuite le terme plus générique de *patron d'indexation*. Ce terme, tout en mettant l'accent sur l'assistance à l'indexation, garde bien une référence à un aspect graphique et relationnel, *via* la notion de *patron*.

– étaient généralement repérées spatialement par rapport à un *lieu* (une maison, une pièce) défini par un degré d'ouverture sur l'extérieur ainsi que par sa fonction sociale.

On peut à partir de ces besoins concevoir un patron d'indexation adapté au point de vue concerné. Le graphe de la figure 3.10 montre ce patron, sous la forme d'un graphe conceptuel dont les sommets sont des individus génériques.

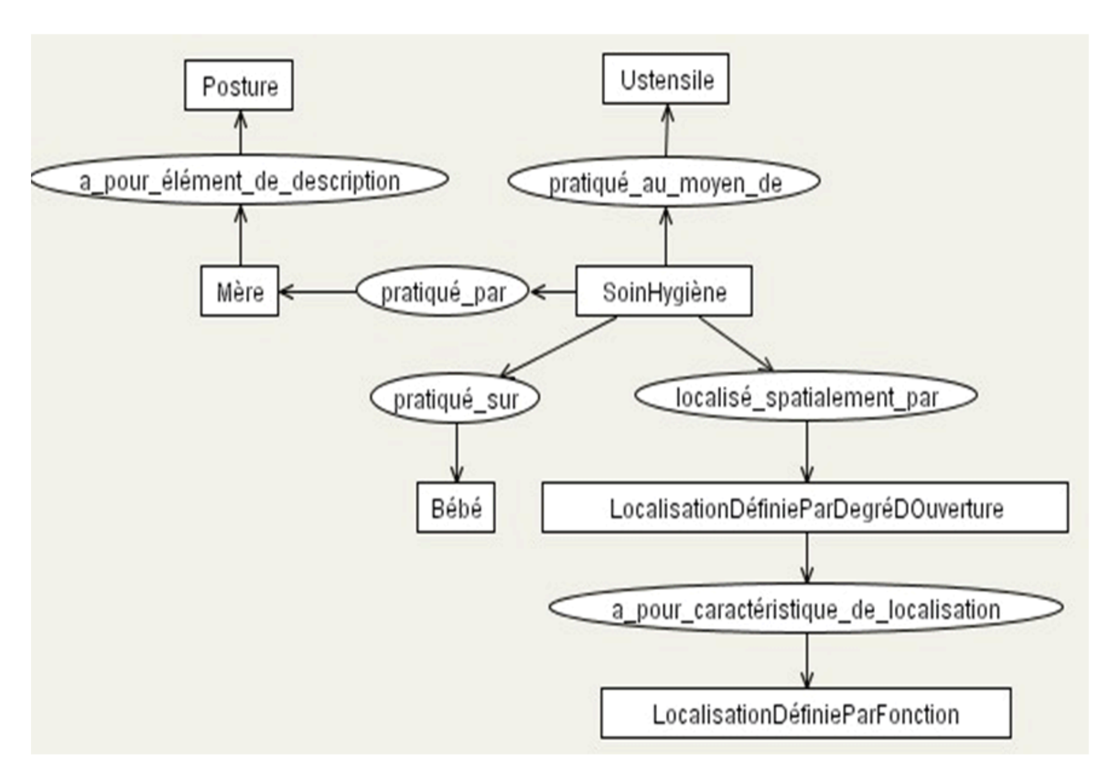


Fig. 3.10 Un patron d'indexation pour la description de soins sur les enfants

Ce graphe sera ensuite modifié en fonction des cas rencontrés et les éléments d'information reconnus pertinents : l'utilisateur peut en particulier :

1. modifier les concepts et relations, par spécialisation ou généralisation,
2. en ajouter d'autres, ou bien en retirer,
3. (ré-)instancier les concepts par des individus nommés.

Toutes ces modifications devront évidemment être en accord avec les contraintes formelles introduites dans l'ontologie, en particulier celles qui sont induites par la donnée des domaines des relations. Le graphe de la figure 3.11 présente un exemple d'index réalisé avec le patron d'indexation de la figure 3.10.

Dans le cas de cet index, on a rajouté une deuxième action, et les deux actions s'inscrivent bien dans un même cadre descriptif, même si la seconde (la *balnéation*) est décrite de manière moins complète que la seconde. Néanmoins, cet index est toujours une spécialisation du patron d'indexation : suivant la sémantique formelle des GC, son interprétation logique implique en effet celle du patron. Mais il peut en être tout autrement : par exemple, pour un segment montrant une mère rentrant le nombril de son nouveau-né, on trouvera l'index de la figure 3.12 qui ne spécialise pas le graphe patron : `Nouveau-né` ne spécialise pas `Bébé`, `SoinAEnfant` est une généralisation de `SoinHygiene` – l'indexeur n'a sûrement pas jugé que l'opération montrée relevait des actes

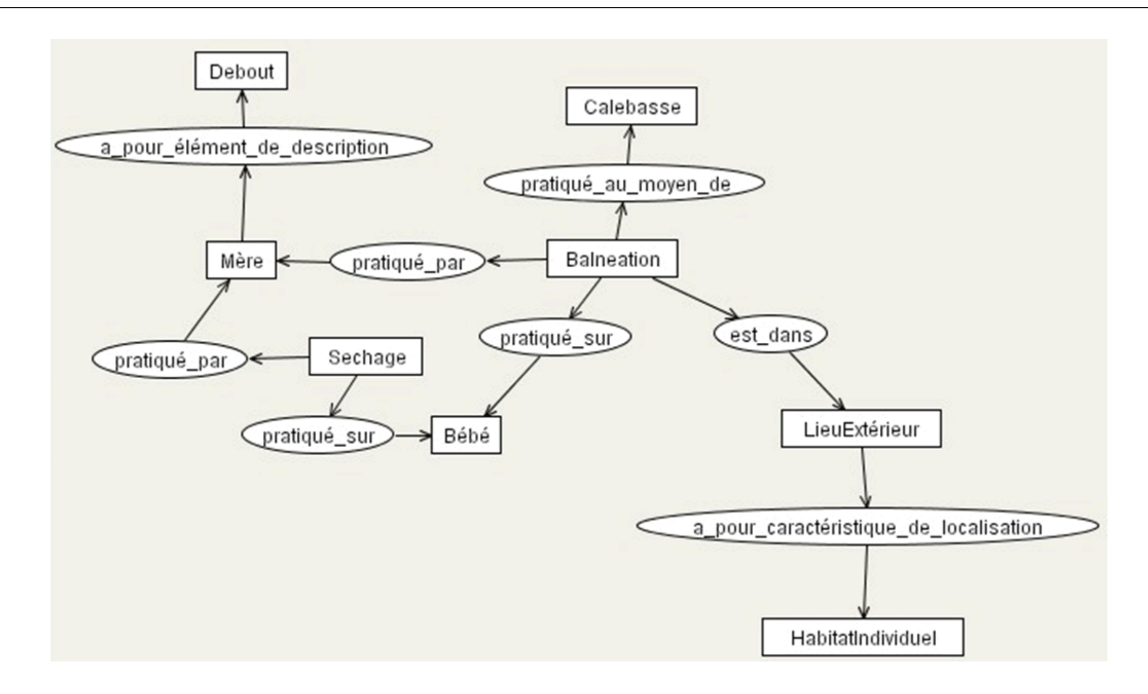


FIG. 3.11 Spécialisation d'un patron d'indexation en un index

d'hygiène usuels – et l'action est décrite comme étant pratiquée sur le nombril du nouveau-né, et non sur celui-ci considéré dans son ensemble, ce qui nous place dans une configuration différente de celle du graphe patron. On a bien ici une illustration de la pertinence (puisque la structure est tout de même très proche, et les concepts, assez peu éloignés dans la hiérarchie de l'ontologie) mais aussi de la souplesse de cette approche.

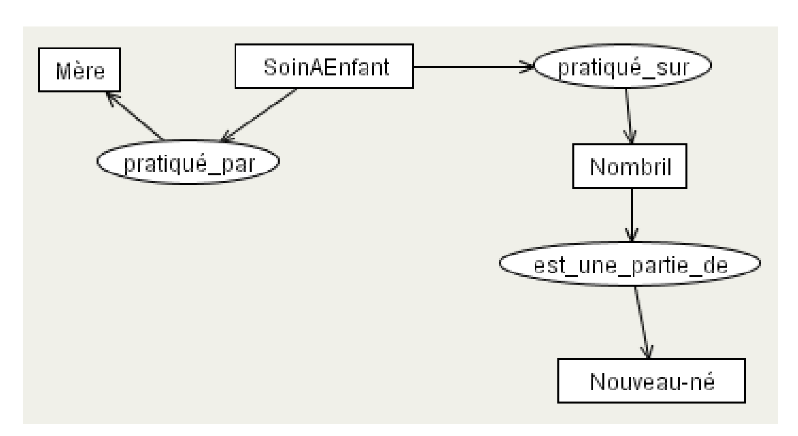


FIG. 3.12 Un index ne spécialisant pas logiquement le graphe patron

Pour synthétiser la notion de patron d'indexation, on peut proposer la définition suivante :

> Un patron d'indexation ontologique est une construction relationnelle adaptable, qui présente, dans un contexte d'indexation typique, l'articulation entre les concepts et les relations ontologiques les plus caractéristiques d'une application. Ce patron sert de point de départ au processus d'indexation, qui consiste à modifier ses éléments et sa structure de manière à rendre compte d'un contenu documentaire donné.

On sait alors quels sont la nature et l'usage de tels patrons. Cependant, plus de précisions sont nécessaires, si l'on veut en particulier cerner le statut du patron et des notions qu'il mobilise, par rapport aux autres ressources ontologiques.

Précisions sur les patrons d'indexation ontologiques

Tout d'abord, il est indispensable de remarquer la complémentarité de l'ontologie et du patron d'indexation. En effet, le patron d'indexation constitue en quelque sorte un mode d'emploi du langage d'indexation qui, comme nous l'avons montré dans la chapitre 2, repose sur l'ontologie de l'application considérée. A ce titre, on ne peut envisager le patron sans l'ontologie : celle-ci clarifie la signification des concepts et relations qu'il utilise, et, par la donnée des notions qui leur sont proches et des contraintes formelles qui s'y appliquent, indique quelles sont les évolutions légitimes qu'il peut subir. De même, dans le cadre d'une application précise, le patron d'indexation est indispensable à l'ontologie : si celle-ci, en fournissant deux interprétations, métier et formelle, donne à ses notions une signification claire dans un contexte d'usage, elle ne donne pas ce contexte d'usage. Le patron, lui, peut être considéré comme la formalisation d'un tel contexte, puisqu'il donne un exemple d'utilisation, précis et explicite, des notions de l'ontologie.

Par exemple, les définitions formelles telles que celles introduites en logique de descriptions, si elles introduisent des motifs qui peuvent à première vue paraître semblables à celles des patrons, n'ont pas le même emploi. Si on sait que l'interview d'un expert est une interview auquel participe au moins un expert, cela ne donne pas d'indication sur la manière d'employer précisément ce concept. On pourrait par exemple croire qu'il est souhaitable d'introduire la description d'un expert pour chaque occurrence, alors que l'information réellement pertinente pour l'application envisagée sera la donnée de ce que l'interview contribue à clarifier, pour reprendre un des points de vue applicatifs introduits dans Opales. Les définitions des concepts et relations donnent les moyens à l'utilisateur et au système de comprendre et d'exploiter les éléments des index, et par là même de faciliter leur emploi lors de la création des index, mais elles ne prescrivent pas réellement de configuration pertinente pour l'application. D'ailleurs, rattachés de fait aux concepts et au relations qu'ils contribuent à définir, ces axiomes ne peuvent clarifier plus que la structure du contexte local[10] de ces notions. Alors que le patron d'indexation a lui vocation à représenter le contexte global d'emploi des notions de l'ontologie. Il n'y a donc pas de contradiction ni de redondance entre ces deux ressources.

Cela nous amène à la discussion sur la quantité de patrons nécessaire pour un seul et même point de vue applicatif. De fait, cela dépend de la précision de l'application envisagée : certaines, comme les points de vue retenus pour Opales, peuvent très bien ne comporter qu'un patron. D'autres, comme c'est le cas dans le projet Menelas [ZC94], proposent une multitude de structures assimilables à des patrons, dans la mesure où nombre des concepts de l'ontologie parmi les plus importants pour le domaine abordé (celui des pathologies coronariennes et des opérations chirurgicales afférentes) sont accompagnés d'un graphe conceptuel introduit comme *modèle de connaissances* du concept. Celui-ci, outre des informations que l'on peut interpréter comme définitoires, comporte des éléments qui sont plus de nature prescriptive, et insèrent le concept dans un réseau relationnel complexe. Par exemple, le concept d'être humain (cf. code 3.1) est associé à un graphe qui énumère certaines des caractéristiques essentielles d'une personne – comme le fait d'avoir un *âge* qui est caractérisé par une *quantité* attachée à une *unité temporelle* – mais aussi certains attributs optionnels. Ainsi, l'âge pourra être associé à une *interprétation qualitative*, ou

[10] Si on utilise le vocabulaire des graphes, il s'agit des nœuds qui sont rattachés directement à celui de la notion considérée.

bien la personne pourra jouer un *rôle* propre au domaine, comme *patient* ou *médecin*, ou être impliquée activement dans un *événement d'origine sociale*.

```
[human_being: _x]-
    (attr)-->[sex ^nec]--(val_qual)-->[sex_val ^nec]
    (attr)-->[age^nec]--(val_quant)-->[quantitative_val^nec]-
        (reference_unit)-->[temporal_duration^nec] ; 30 years old
        (categorised_as)-->[general_category]% ; old, young
    (defines_cultural_function)
        -->[medical_subfunction]--(cultural_role)-->[patient]
    (defines_cultural_function)
        -->[medical_subfunction]--(cultural_role)-->[physician]
    (participate_to)-->[social_event]
```

CODE 3.1 – Extrait du modèle de connaissances du concept `human_being` de l'ontologie MENELAS

Pour MENELAS, il n'y a donc pas de graphe patron reflétant un besoin unifié d'indexation, mais une quantité importante de prescriptions de type patron, chacune centrée sur un type d'entité privilégié. Cela est très lié au fait que ces ressources servent en fait à guider une indexation automatique, qui à partir des manifestations textuelles des concepts isolés essaient de construire un réseau complexe, suivant justement les patrons associés à chacun de ces concepts. Une telle forme de préconisation, si elle est légitime dans l'absolu, présente beaucoup de désavantages pour une indexation manuelle : il faut en effet avoir déjà choisi le concept central de la description, parmi les nombreux candidats dont on peut supposer qu'ils sont tous d'un intérêt à peu près équivalent pour l'application envisagée. L'indexeur se retrouve donc livré seul face au vaste répertoire des concepts de l'ontologie, ce qui ne favorisera pas son travail d'interprétation et de formulation. En fait, cette situation est analogue à ce que nous avons observé, dans un cadre documentaire classique, pour le cas de l'INA. Nous avons bien des grilles d'indexation qui privilégient des choix de descripteurs, mais celles-ci sont en très grand nombre, les sujets considérés comme pertinents au point d'être ainsi formalisés se comptant par dizaines, voire par centaines[11]. Une telle confusion reflète évidemment une application dont les objectifs sont très généraux, et donc difficilement élicitables.

En pratique, plus on voudra prescrire finement une indexation, plus réduite devra être la quantité de patrons. Un tel objectif, s'il peut sembler naïf, n'est pas non plus contradictoire. Si l'ontologie est extrêmement complexe à comprendre et utiliser, c'est qu'elle vient avec des spécifications qui sont également complexes. Et qu'on ne peut avoir atteint un tel niveau de complexité sans avoir précisé exactement les interprétations des descripteurs, et, logiquement, les descriptions à effectuer. De fait, comme on le verra dans la section suivante, l'emploi des patrons d'indexation peut être relié au fonctionnement inférentiel d'un système à base de connaissances. En tout cas, il s'agit d'atteindre un compromis entre le nombre de patrons et leur simplicité qui ne pourra être encore une fois déterminé qu'au cas par cas, en fonction du contexte applicatif de l'indexation, et notamment des besoins en termes de richesse des notions et des situations à décrire.

[11]Un décompte des ressources que nous avons collectées au cours de notre travail fait état de cent cinquante de ces grilles, de la Coupe de France de basket aux accidents de centrales nucléaires, en passant par la vente du muguet le 1er mai. Mais ce chiffre est inférieur à la réalité, puisque nous n'avons eu accès qu'à une partie de ces grilles, et que, d'autre part, ce processus de collection s'inscrit dans le cadre d'un effort de capitalisation des connaissances des documentalistes qui est assez récent et toujours en cours.

Se pose ensuite la question de la spécificité des notions mobilisées dans un patron : plus concrètement, à quel niveau se place-t-on dans la hiérarchie des concepts ou des relations d'une ontologie ? Là encore, il n'y a pas de réponse générale, la *typicité* par rapport à l'application évoquée dans notre définition restant très subjective. Mais on peut donner quelques réflexions résultant de nos observations. Tout d'abord, il évident que les patrons ne sont pas donnés au niveau le plus haut qui soit disponible dans l'ontologie et qui valide les contraintes qui y sont spécifiées. Dans une ontologie complète[12], on peut en effet distinguer trois niveaux[13] :

- le niveau le plus haut (*upper-level*) introduit les types les plus génériques permettant de catégoriser les entités du monde : on y trouvera ainsi des concepts comme `ObjetSpatial` (les entités tangibles du monde), `ObjetTemporel` (les événements ou processus) ou bien encore des `ElementDeDescription` (les propriétés que l'on attribue à des individus) ;
- le niveau le plus spécifique – [Bac04a] le qualifie comme *parataxique* – est celui qui présente les notions spécifiques à l'application considérée. Dans une application d'indexation, ce seront les concepts et les relations effectivement utilisés pour catégoriser le plus finement possible les entités du monde introduites dans les descriptions. Dans le domaine de la petite enfance, on trouvera par exemple les concepts `MèreAdoptive`, `MèreNourrissière`, `MèreCélibataire`...
- enfin, il existe un niveau intermédiaire. Celui-ci est spécifique à un domaine mais introduit des catégories plus générales qui structurent ce domaine – [Bac04a] utilise l'expression de *niveau structurant*. Ces notions sont celles qui sont utilisées pour définir[14] et regrouper les notions plus spécifiques du niveau de l'application.

L'hypothèse raisonnable[15] que nous faisons est que les notions utilisées dans le patron d'indexation relèvent de ce dernier niveau. Cette hypothèse repose sur nos expérimentations, notamment pour Opales, mais aussi sur des considérations pratiques. Les patrons n'ont pas en effet à rendre compte directement des lois contraignant l'usage des concepts et relations dans l'ontologie. Par exemple, pour notre point de vue de la petite enfance, il serait improductif de présenter un patron articulant des `actions` impliquant, en tant que `participants` génériques, des `Objets Physiques` ou des personnes qui seraient `caractérisés par` des `Eléments de Description`, notions comptant parmi les plus abstraites de l'ontologie développée pour l'occasion. Plus la notion donnée dans le patron est générique, plus l'indexeur aura d'effort pour adapter le patron d'indexation à un besoin de description concret.

[12]On fait par là allusion à une ontologie qui contient tous les concepts qui sont nécessaires à l'application, tout en étant *close pour la définition*, ce qui implique qu'elle contient tous les concepts – y compris le concept applicable à toute entité – demandés pour définir ceux-ci. Traditionnellement, on préfère présenter les concepts dans des *modules* dont les niveaux de généralité sont différents, mais notre approche par définition différentielle et le besoin de situer les notions des patrons sur une échelle « absolue » nous poussent ici à considérer sous un angle unificateur les ressources ontologiques utilisées.

[13]Il en existe un quatrième, le niveau *formel*, qui s'attache à la description des modalités d'existence des types introduits dans les ontologies de haut niveau ou de domaine. Par exemple, en s'intéressant au comportement, dans le temps et dans l'espace des possibles, d'un lien unissant une instance à son type. Nous reverrons ces considérations dans la section 4.2.2.

[14]On peut retrouver ces notions dans les définitions formelles des ontologies, mais de fait nous visons surtout ici les définitions langagières, plus proches des mécanismes cognitifs supposés être en jeu chez les experts du domaine. Il serait ainsi tout à fait envisageable et naturel, dans le domaine de la petite enfance, de paraphraser « mère nourrissière » par « mère reconnue comme telle car ayant contribué à l'alimentation de ses "enfants" », expression qui met bien en valeur la notion *structurante* de mère dans ce cas.

[15]Cette hypothèse ne peut cependant être vérifiée que si l'on dispose effectivement de « strates » aisément identifiables dans l'ontologie étudiée, ce qui suppose qu'elle a été conçue suivant des principes précis. Dans la pratique, une telle situation n'est absolument pas garantie. Cela est particulièrement vrai en ce qui concerne les relations, qui sont souvent introduites dans des hiérarchies de profondeur moindre que celles des concepts, parfois même sous la forme de listes non hiérarchisées.

Il ne faut pas non plus aller vers le trop spécifique : le patron n'a pas vocation à être un pur exemple – illustratif et non prescriptif – à partir duquel l'indexeur devrait extrapoler un besoin d'indexation. En particulier, il semble absurde de proposer un patron qui introduise des instances nommées pour les concepts présents, sauf en cas de besoin extrêmement précis, comme une application relative à un objet particulier (personne ou lieu, par exemple) du domaine. Des instances génériques seront généralement plus adaptées à ce que le patron d'indexation est censé exprimer : la manière dont les descriptions pertinentes sont généralement formulées.

Enfin, on peut s'interroger sur la complexité du patron d'indexation et des index qui sont obtenus à partir de celui-ci, à la suite des opérations de modifications que nous avons introduites en page 99. Ici également on ne peut préjuger de ce que seront les besoins des points de vues applicatifs à venir, et pour chacun des points de vue, de tous les contenus documentaires à interpréter et représenter. Idéalement, un patron d'indexation doit être plutôt simple, pour faciliter son appropriation et son utilisation par les indexeurs. Par contre, il ne doit pas être simpliste, et favoriser la création d'index suffisamment riches au regard des besoins applicatifs.

Par exemple, dans le point de vue d'OPALES relatif à la description des documentaires relatifs à la géographie, nous avons dégagé un patron en apparence simple, composé d'un nombre restreint de concepts (cf. figure 3.13).

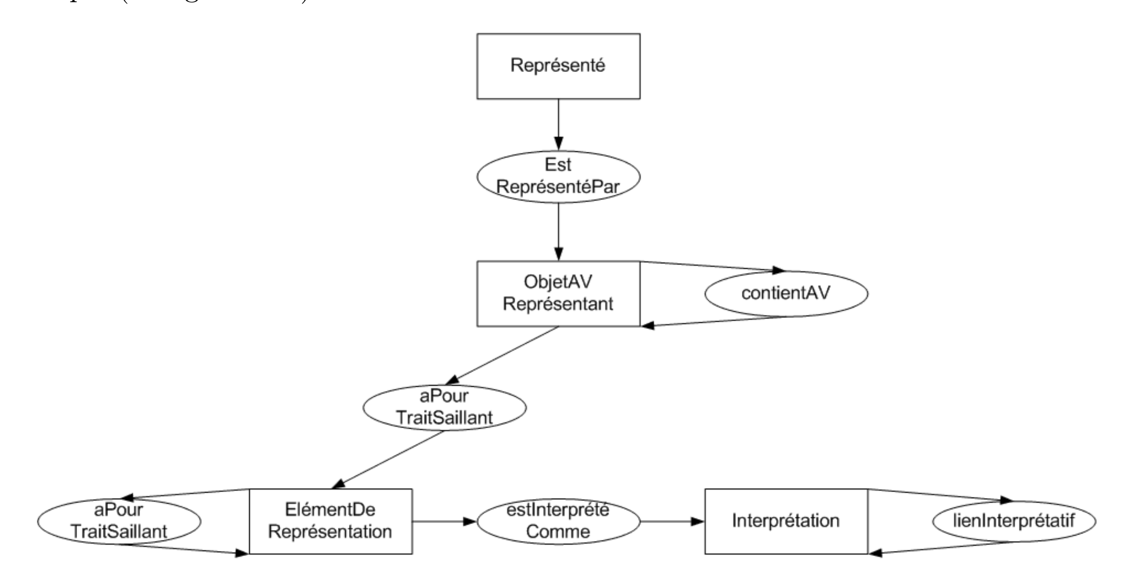


FIG. 3.13 Le patron d'indexation du point de vue applicatif « eau et audiovisuel »

Ces concepts sont néanmoins liés par des relations qui en plus de structurer l'index de façon complexe, introduisent une certaine forme de récursivité. Les objets audiovisuels (images, séquences) introduits peuvent en effet faire l'objet d'une décomposition, tout comme les éléments qui ont été retenus comme intéressants du point de vue de l'analyse de l'image (les « traits saillants ») peuvent à leur tour être caractérisés par certains sous-éléments tout aussi intéressants mais d'un grain plus fin. Enfin, les interprétations induites par l'emploi de ces éléments peuvent elles-même être insérées dans un réseau qui les situe les unes par rapport aux autres (relation d'opposition, de clarification. . .). A l'arrivée, les indexeurs se sont étonnés de la complexité que leurs index atteignaient, alors qu'ils n'avaient pas l'impression de s'être écartés du schéma qui leur était proposé. La figure 3.14 donne un exemple caractéristique de la complexité que l'on peut ainsi atteindre, par l'application d'une quantité relativement limitée – et donc gérable

par l'utilisateur – d'opération de décomposition des objets introduits et d'ajouts de nouveaux extraits de patrons. En l'occurrence, la séquence indexée a été décomposée en deux sous-objets, chacun étant décrit à son tour selon les indications apportées par le patron.

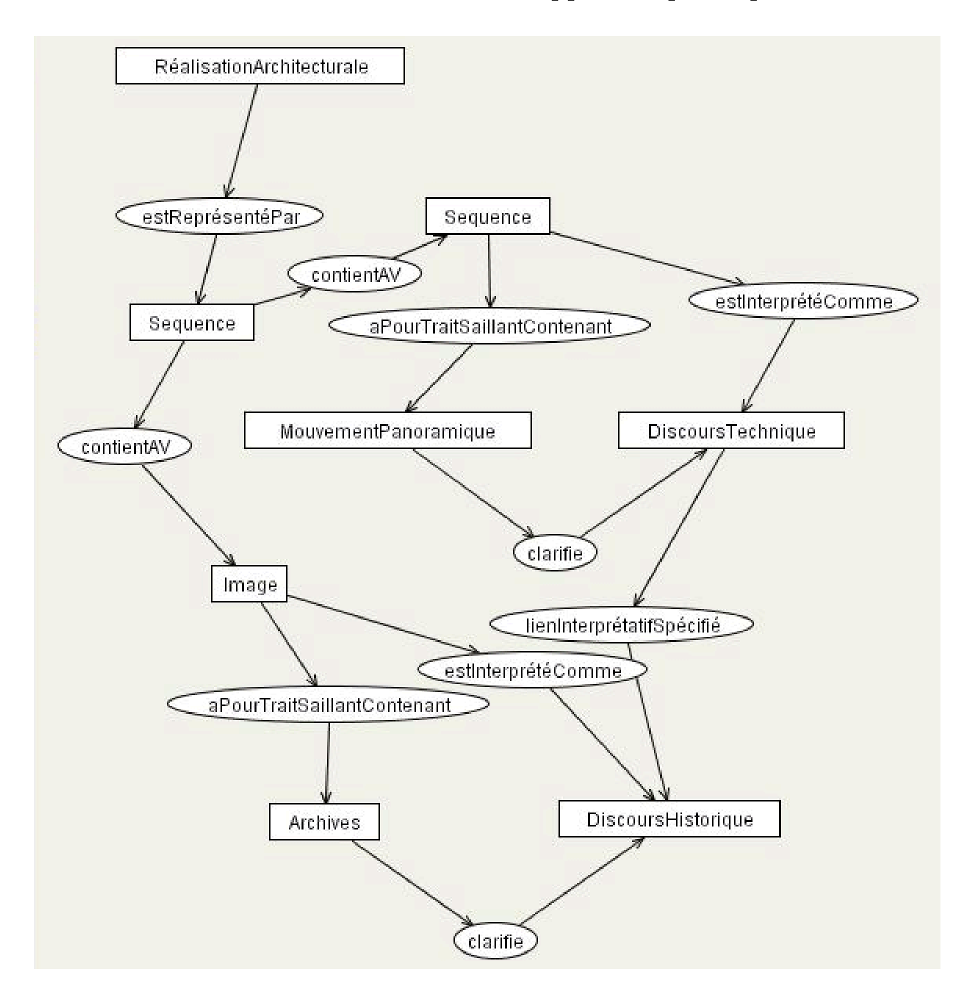


Fig. 3.14 Un index obtenu à partir du patron « eau et audiovisuel »

Grâce aux patrons d'indexation, un indexeur dispose d'une structure de connaissances accessible, explicite, globalement pertinente pour le point de vue applicatif retenu. Les patrons sont un moyen de prescrire une indexation standard, ce qui augmente le rendement de l'ensemble du système. Et l'on prend en compte correctement et de manière flexible les besoins liés au contexte applicatif, ce qui améliore la pertinence du système.

De fait, dans une analyse de la conception et de l'usage des ontologies dans les systèmes d'information, [UG96] a introduit les *questions de compétence* comme moyen de spécifier et d'évaluer les connaissances qu'elles sont amenées à contenir. En première analyse, les ontologies doivent évidemment fournir le vocabulaire et les connaissances de raisonnement nécessaires au bon fonctionnement du système d'indexation. Mais ce bon fonctionnement, dans le cadre d'une indexation demandant une certaine forme de contrôle éditorial, implique de savoir comment utiliser le vocabulaire ontologique pour créer des index. Les patrons d'indexation apportent une réponse tangible à cette première question de compétence : en complément de l'ontologie, ils

prescrivent une manière de produire facilement des indexations pertinentes et cohérentes.

3.3.3 Utilisation des patrons d'indexation et raisonnement

L'intérêt des patrons va cependant au-delà de la simple création des index. Dans un SBC d'indexation et de recherche, en effet, on utilise des mécanismes d'inférence formelle dont l'objectif est de répondre à une deuxième question de compétence : comment trouver des résultats pour les requêtes posées aux systèmes ? On a vu au chapitre 2 que l'enjeu est alors de rapprocher les requêtes des connaissances dont on dispose dans la BC, c'est-à-dire des index. Le point précis que nous voulons souligner ici est que les connaissances de raisonnement « proches » des patrons d'indexation sont particulièrement importantes dans ce cadre.

L'emploi de ces patrons est en effet lié à une certaine forme de standardisation de l'indexation, standardisation qui s'inscrit bien dans le cadre déjà abordé de la continuité sémantique, puisqu'en rapprochant les index d'un modèle pertinent pour l'application visée on augmente les chances pour ces index d'être correctement interprétés et exploités dans ce contexte applicatif. De fait, l'exploitation des index par un moteur d'inférence utilisant des connaissances ontologiques peut bénéficier du recours à des patrons. En effet, les configurations et les notions que ceux-ci présentent sont supposés proches des index produits, ce qui autorise à les considérer comme des structures *pivôts* auxquelles on peut comparer les requêtes et les descriptions de manière à faciliter leur rapprochement.

La première solution à laquelle on puisse penser est d'introduire les patrons en tant que spécifications des requêtes à poser par les utilisateurs. De la même façon que l'indexeur avait accès à une formulation typique des connaissances du domaine, le chercheur aura un aperçu de la forme canonique des index de la base qu'il s'apprête à interroger. Ce chercheur pourra ensuite adapter cette forme à son besoin spécifique : dans le cadre d'un moteur de recherche exploitant les graphes conceptuels, il créera ainsi un graphe représentant sa requête et le système se chargera de trouver les index qui spécialisent (et donc impliquent au sens logique) cette requête. Dans cette approche, le patron d'indexation sert de référentiel commun à la création des index et à celle des requêtes ; nous avons donc là un premier moyen de réduire la distance entre ces éléments.

Cependant, le processus d'adaptation du patron, que ce soit à des documents particuliers au moment de l'indexation ou à des besoins informationnels précis au cours de la recherche, crée une distance indéniable. Cette distance devra alors être comblée grâce aux mécanismes d'inférence exploitant les connaissances de raisonnement formelles de l'ontologie, qui permettent de simuler certaines des reformulations opérées habituellement par les utilisateurs du système documentaire, comme nous avons vu au chapitre 2.

Ainsi, les liens de subsomption formelle qui guident les processus de spécialisation ou de généralisation des concepts et des relations introduits dans le patron peuvent être vus comme des moyens de compenser les effets de ces processus. Par exemple, l'index de la figure 3.11, obtenu à partir de spécialisation et d'ajout successif d'éléments du patron de la figure 3.10, implique toujours logiquement ce patron, du fait des liens de subsomption qui permettent de projeter les concepts et relations de l'un sur ceux de l'autre. L'index pourra donc répondre à une question formulée directement à l'aide du patron, ou alors obtenue par généralisation ou suppression des notions qu'il présente – par exemple, si le chercheur demande des documents où une personne quelconque participe à une action ayant pour objet un bébé.

Les choses deviennent plus délicates lorsque l'utilisateur modifie l'agencement des concepts et des relations du patron. Par exemple, l'index de la figure 3.12 ne spécialise plus le patron d'indexation du point de vue dont il relève. Pour rapprocher les deux, on devra utiliser des connaissances de raisonnement établissant des liens entre des structures de connaissance diffé-

rentes. Dire par exemple que l'on considère, pour notre application, qu'une action pratiquée sur une partie d'un objet ou d'une personne l'est sur cet objet ou cette personne. Les règles de raisonnement s'intéressant prioritairement à la composition des relations conceptuelles s'inscrivent souvent dans une telle optique. Dans les expériences que nous avons menées, on trouve ainsi toute une série de connaissances de raisonnement dont l'objectif est de ramener des connaissances formulées de manière « non standard » à un format plus proche de celui qui sera utilisé pour les index et surtout les requêtes. La figure 3.15 montre des exemples de règles relationnelles issues du domaine applicatif de la petite enfance et ayant cet objectif. Outre celle du « transfert » de la relation de participation d'une partie à un tout, on peut trouver une règle spécifiant qu'une localisation peut être attribuée à une action si on connaît un repérage spatial pour l'un de ses participants. D'une configuration légèrement différente de celle du patron d'indexation, on peut ainsi déduire des informations[16] dont la structure est désormais conforme à ce que l'on attend généralement.

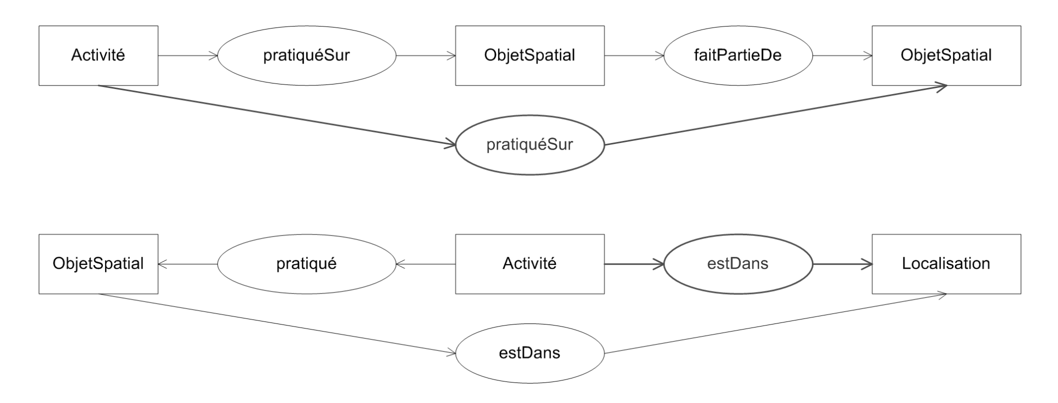


Fig. 3.15 Règles de raisonnement déduisant des connaissances conformes à un patron d'indexation d'une structure de connaissance différente

Les connaissances explicitées par le raisonnement sont représentées en gris.

La réciproque est également possible : à partir des connaissances exprimées dans une structure conforme à celle du patron d'indexation, on peut vouloir expliciter de nouvelles connaissances. On pourra ainsi répondre à des questions dont l'agencement des notions est différent de celui du patron, et dont les besoins ne pourraient donc pas directement être satisfaits par celui-ci. Ainsi, pour continuer dans notre exemple, le patron d'indexation de la petite enfance associe de façon privilégiée une action à sa localisation, et ne préconise pas le repérage des participants. Pourtant, de nombreuses requêtes pourraient exiger une telle information, déductible naturellement de la localisation des activités décrites. Par exemple, si une action est située dans un lieu donné, il est légitime de considérer que ses participants sont également situés dans ce lieu, puisque les activités dans le domaine considéré sont des activités « simples » mettant en jeu des acteurs physiquement présents au même endroit. Il sera alors pertinent de munir l'ontologie d'une règle de raisonnement telle que celle de la figure 3.16.

L'emploi des patrons d'indexation est donc intimement lié à celui des connaissances de raisonnement, même si le niveau auquel ces deux types de ressources sont données est différent. En effet, pour pouvoir s'appliquer au plus grand nombre de cas avec pertinence, les règles d'in-

[16]En fait, il s'agit, comme dans les processus de recherche documentaire traditionnels, de ré-expliciter une information que l'on a gardé implicite par souci d'économie.

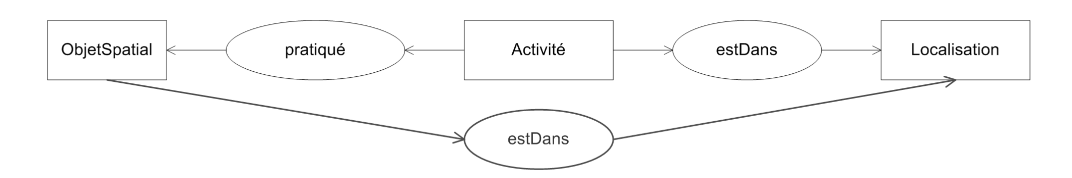


FIG. 3.16 Règle de raisonnement déduisant de nouvelles connaissances à partir de celles d'un patron d'indexation

férences doivent produire les connaissances les plus spécifiques que l'on puisse dériver à partir des hypothèses les plus générales possibles. Ainsi, dans l'exemple de la localisation, on cherche à transférer le type de repérage le plus précis qui soit – `estDans`, `estSur`, ce qui impliquera de créer autant de règles que de type de repérage –, mais en l'appliquant à des entités extrêmement générales – `ObjetSpatial` et `Activité`. L'utilisateur n'a pas en effet à être confronté à ces connaissances, qui sont justement conçues pour faciliter sa tâche de manière transparente, en assouplissant les contraintes imposées par le contrôle éditorial que constitue la donnée des patrons. Les règles ne doivent donc pas cibler exclusivement les concepts et relations structurants du domaine, mais au contraire tenter de compenser des variations par rapport à ces notions qui soient d'une amplitude maximale.

3.4 Conclusion

On a constaté qu'il était nécessaire de compenser à l'aide de propositions techniques ou méthodologiques la complexité apportée dans les processus documentaires par l'utilisation d'outils de représentation et d'exploitation des connaissances formalisées. Pour cela, il faut essayer de se rattacher systématiquement aux usages rencontrés dans l'application visée par le système documentaire. Ceci concerne tout aussi bien la compréhension des notions employées que la façon dont celles-ci sont utilisées pour créer des descriptions des contenus documentaires à même de satisfaire les besoins auxquels le système est supposé répondre – en l'occurrence, ceux de la recherche documentaire.

Lors de l'ouverture de cette seconde partie, nous avons donné une liste de points méthodologiques à la lumière desquels il est possible de récapituler les contributions de ce chapitre. Tout d'abord, **M1** mettait en avant l'impératif de faciliter la compréhension des termes de la description. Nous avons vu que pour cela il est naturel de se tourner vers la langue. On peut en particulier structurer les significations non formelles des intitulés langagiers attachés aux concepts et aux relations par une normalisation sémantique s'appuyant sur des principes de similarité et de différence au sein d'un réseau de significations. Reprenant ces propositions de Bruno Bachimont, que nous avons mises en œuvre au cours de cette thèse, nous avons détaillé en quoi consistait une *ontologie différentielle* et comment on pouvait présenter à l'utilisateur des concepts et des relations dont l'interprétation serait immédiate.

Ensuite, le point **M3** insistait sur la difficulté de créer des index à base de connaissances complexes, qui puisse faire bénéficier le système d'indexation et de recherche des traitements riches rendus possibles par la formalisation sous-jacente aux systèmes ontologiques. Nous démarquant des techniques déjà existantes, que ce soit dans les systèmes classiques ou ontologiques, nous avons détaillé comment une solution avancée pour OPALES, les *patrons d'indexation*, peut offrir un cadre propice à un contrôle éditorial qui mette en lumière des configurations pertinentes pour l'application visée, tout en gardant un niveau de souplesse suffisant pour ne pas rebuter

l'indexeur. En fait, les patrons d'indexation sont à utiliser en conjonction avec des connaissances de raisonnement formelles qui permettent justement de faire la jonction avec les différentes déclinaisons de son contenu, que ces déclinaisons soient présentes dans les requêtes ou les index. On a alors réduit doublement la charge cognitive reposant sur l'indexeur : on lui propose une structure canonique pertinente, et il dispose d'une marge de manœuvre significative pour encoder ses propres interprétations sans pour autant les rendre inaccessibles pour les autres utilisateurs.

Il faut noter qu'on est alors dans la portée du point **M2**. L'application des connaissances de raisonnements a à présent une légitimation immédiate vis-à-vis des processus de description – dont elle assouplit les contraintes éditoriales ou formelles – et de recherche – cet assouplissement va de pair avec des capacités de recherche qui ne sont aucunement amoindries. Elle n'est plus un obstacle à la compréhension des SBC d'indexation, et donc à leur acceptation par des utilisateurs non experts en représentation des connaissances.

De fait, les propositions de ce chapitre ont des répercussions importantes sur la manière dont les ressources ontologiques devront être conçues. Par exemple, les patrons d'indexation doivent être pris en compte avec soin au moment de la spécification du vocabulaire ontologique et des connaissances de raisonnement afférentes. Ils présentent les concepts et les relations reconnus comme pertinents pour l'application, et une partie de l'activité de spécification ontologique devra chercher à anticiper les variations dont il pourront faire l'objet. C'est alors qu'on devra créer les connaissances de raisonnement qui tenteront de compenser ceux de ces écarts qui seront reconnus comme légitimes au regard des besoins applicatifs.

On doit donc se reposer la question, déjà évoquée dans la présentation des points méthodologiques mentionnés ci-dessus, de la rationalisation du processus de conception d'ontologies formalisées correspondants à des besoins précis. En d'autres termes, il faut repenser la conception d'ontologies – notre problématique *disciplinaire* – pour tenir compte des contraintes d'usages – notre problématique métier. Il va s'agir en particulier d'intégrer à ce processus deux éléments méthodologiques – ontologie différentielle, patrons d'indexation – qui sont nécessaires à une *situation* correcte des artefacts construits dans les usages observés dans l'application visée. Et de voir comment cette intégration peut se faire de manière à favoriser la *réutilisation* d'éléments déjà existants, le *partage* des éléments conçus, et ainsi que l'*interopérabilité* avec d'autres ontologies. Ceci pourrait grandement faciliter et légitimer le travail des spécialistes en RC chargés de la conception des ontologies, tant parce qu'ils auront moins à concevoir eux-mêmes – en réutilisant des notions abstraites consensuelles adaptables à leur application – que parce que ce qu'il feront sera à son tour consensuel et adaptable pour la conception d'autres ontologies, ou réutilisable dans d'autres systèmes.

Chapitre 4

Faciliter la conception d'ontologies pour l'indexation sémantique

4.1 Introduction

Dans les chapitres qui ont précédé, on a vu comment l'utilisation de SBC utilisant des ontologies était en mesure d'apporter des solutions novatrices aux problèmes de l'indexation et de la recherche documentaires. Complexes du point de vue de leur utilisation, ces ontologies peuvent néanmoins s'insérer dans les usages et les compréhensions rencontrés dans les applications concernées. Il faut pour cela créer les ressources – spécifications formelles et informelles, patrons d'indexation – dont on vient de voir qu'elles étaient nécessaires à une utilisation et un fonctionnement pertinents du système dans ce cadre applicatif.

Le problème est que de telles ressources ne sont pas faciles à construire, même si l'on suppose données les connaissances du domaine qui permettront de le faire– par la collaboration d'un expert, ou l'expertise du concepteur de l'ontologie lui-même. En effet, on a vu que pour bénéficier des pleines possibilités d'une approche ontologique, et en particulier de ce qu'offrent les mécanismes de raisonnement en termes de production et de contrôle des connaissances, il faut des ontologies complexes, dont les concepts et les relations seront rattachés à des connaissances de raisonnement.

De telles ontologies sont lourdes à concevoir : l'exemple de **Cyc** [Len95], ontologie « universelle » dont la mise au point s'étend sur des années, est là pour en donner une idée. Si l'immense majorité des ontologies à développer n'a pas pour objectif l'axiomatisation complète de toutes les catégories abstraites qui structurent la pensée humaine, il n'en reste pas moins que développer des spécifications formelles, même dans un domaine limité, n'est pas chose aisée. Pour preuve, le petit nombre d'ontologies dont le niveau de complexité est élevé[1]. [TV03], qui a cherché à catégoriser les ontologies d'un répertoire public[2] utilisant un langage dédié à la spécifications d'ontologies en LD, trouve que le plus gros contingent regroupe en fait des vocabulaires n'utilisant que les axiomes de spécialisation, et introduisant très peu de relations de domaine. De son côté, [HNM02], analysant un corpus d'ontologies plus anciennes mais utilisant un langage plus riche, révèle que seules 10% d'entre elles contiennent des axiomes complexes…

Depuis ces dernières années, on a pu observer un regain d'intérêt pour des ontologies plus

[1] On utilise pour qualifier de telles ontologies le qualificatif de « lourdes » (*heavyweight*). On a une opposition entre ces ontologies lourdes et les ontologies « légères » (*lightweight*) qui en général se limitent à l'introduction de simples hiérarchies de concepts.

[2] `http://www.daml.org/ontologies`

élaborées[3], mais celui-ci n'a pas vraiment coïncidé avec un accroissement de la circulation de tels artefacts, contrairement à ce que l'on pourrait espérer. Créer une ontologie complexe requiert des efforts importants, et seules les initiatives de recherches publiques (DOLCE, GALEN) sont propices à la mise à disposition relativement libre d'une telle valeur ajoutée.

La tâche ne peut donc être que plus ardue si, comme nous, on veut prendre en compte correctement des usages précis, et apporter des solutions adaptées aux impératifs d'accessibilité et de pertinence qui découlent d'un tel choix. Nos propositions en matière d'utilisation des ontologies ont nécessairement un impact sur leur construction. Il faut que les ontologies, en tout cas dans le cadre de l'indexation[4], soient conçues de manière à prescrire des spécifications en accord avec les interprétations naturelles du domaine, et leur vocabulaire doit s'insérer dans des usages descriptifs précis, formalisés par les patrons d'indexation. Ceci précise en quelque sorte des objectifs complémentaires à ceux qui sont traditionnellement mentionnés – définition d'un vocabulaire, encodage de celui-ci selon un méta-langage de représentation formelle. . . .

L'enjeu de ce chapitre est donc de revenir à la problématique disciplinaire que nous avons évoquée dans notre introduction générale. Comment les considérations applicatives que nous avons développées jusqu'à présent peuvent-elles être insérées dans le processus de conception des ontologies ? Peut-on proposer des outils concrets à même d'assister les concepteurs d'ontologies dans leurs travaux ? A ce stade de notre réflexion, on sent bien que la détermination d'objectifs précis pour la conception des ontologies peut être un moyen d'améliorer et de faciliter le processus de modélisation formelle des connaissances. Peut-on formaliser l'influence de nos propositions sur ce processus de sorte qu'elle le guident et lui apporte une légitimité immédiatement identifiable, au lieu de le complexifier sans contrepartie évidente pour son auteur ?

Cela implique en particulier de réflechir au processus de formalisation. Comment articuler, d'une part, la détermination d'une interprétation formelle pour les concepts et les relations de l'ontologie, visant des calculs symboliques dans le cadre du SBC, et, d'autre part, l'expression naturelle des connaissances ? Par exemple, si l'on recherche un processus d'indexation qui autorise les utilisateurs à employer certaines formes d'économie conceptuelle pertinentes pour l'application, il ne faut pas que cela se passe au détriment de la cohérence des spécifications formelles précises que l'on attribuerait aux notions introduites dans l'ontologie. On ne doit pas perdre de vue que les ontologies ont vocation à établir une certaine forme de consensus, aussi nécessaire à leur partage au sein d'une communauté que leur pertinence par rapport à un système donné.

Au cours de ce chapitre, nous allons donc discuter de principes de conception qui ont pour objectif d'obtenir des ressources ontologiques – concepts, relations, spécifications formelles de raisonnement, patrons d'indexation – qui soient tout à la fois :
- *complètes* : on veut créer des spécifications formelles précises, qui autorisent les raisonnement riches requis par l'application ;

[3]Dans la cadre du web sémantique, notamment, l'introduction de méta-langages de représentation complexes, comme OWL a accompagné la reconnaissance du besoin des raisonnements dans les applications qui dépassent le statut de « jouets ». Et l'on peut noter une montée en puissance des efforts de recherche visant à la modélisation de connaissances de raisonnement toujours plus complexes [HPB+04] et à la réutilisation d'approches et d'outils relevant d'une longue tradition de raisonnement logique en intelligence artificielle.

[4]Nous rappelons que nous nous sommes placé dans un tel cadre, qui apporte les contraintes que nous nous sommes attaché à prendre en compte depuis le début de ce travail de thèse. Nous soulèverons de fait en conclusion de cette thèse le problème de la généricité des solutions retenues, en particulier en ce qui concernent celles évoquées dans ce chapitre. Il ne semble par exemple pas tout à fait irrationnel qu'une ontologie, même si elle n'est pas supposée être employée dans un contexte d'indexation, ne perdra rien à être conçue en accord avec des compréhensions métier. . .

– *cohérentes* : les spécifications doivent respecter des critères qui leur garantissent un niveau satisfaisant de crédibilité, et donc de partageabilité, tout en restant en accord avec les spécifications applicatives découlant notamment des contraintes de compréhension et d'utilisation.

Comme cela a été évoqué dans maints travaux relatifs au sujet ([GFC04, Bac04a]), il n'est peut-être pas possible d'envisager *le* guide méthodologique définitif en ce qui concerne la conception des ontologies, une activité qui s'apparente plus à un savoir-faire d'ingénierie qu'à une science. Mais cela ne doit pas nous empêcher de proposer des réflexions solides quant à des possibilités concrètes d'amélioration de ce processus.

Nous allons dans un premier temps observer les solutions qui ont déjà été avancées pour faciliter la tâche de conception des ontologies et en améliorer le résultat. Dans la lignée des travaux que nous avons effectués au début de notre thèse [BIT02, TI02], nous montrerons qu'en la matière, s'il existe de nombreuses propositions de méthodes ou bien d'outils de conception, peu ciblent les usages autant que nous devons le faire. Et que ces propositions, qui concernent surtout la gestion du cycle de vie des ontologies, l'assistance à la saisie ou le contrôle des spécifications formelles en dehors de toute visée applicative explicite, doivent être complétées par d'autres, guidant le concepteur de manière plus précise. Nous verrons en particulier que le problème des taxonomies est souvent abordé de manière sommaire, alors que cet élément est central pour la compréhension et l'utilisation des artefacts ontologiques, et donc pour leur conception.

Pour résoudre cette difficulté, nous allons nous tourner à nouveau vers les propositions de Bruno Bachimont, qui insèrent les prescriptions interprétatives que nous avons vues au chapitre 3 dans une réflexion globale sur la conception d'ontologies, de la précision des significations linguistiques rencontrées dans un domaine à la création de référentiels opérationnels disponibles pour des SBC concrets. Nous présenterons cette méthode, à laquelle nous avons contribué au travers d'expérimentations significatives et d'une mise en œuvre dans un outil d'édition d'ontologie, **DOE**.

Cette section constitue le premier pas vers une prise en compte efficace de nos considérations applicatives dans le cours de la conception des ontologies. Après le cas de la normalisation sémantique qui se préoccupe en priorité de la substance des index ontologiques – concepts et relations –, vient le cas des patrons d'indexation, tournés vers leur forme. Pour obtenir des résultats cohérents tant d'un point de vue pratique que théorique, nous proposons de reprendre, en la précisant, l'approche des *patrons de conception ontologiques*. Certains auteurs avancent en effet dans leurs travaux des structures génériques, réutilisables au sein d'ontologies de domaine spécialisées. Nous allons montrer que cette vision, telle qu'elle a été introduite, peut être trop rigide au regard de besoins concrets, mais qu'elle peut aisément être adaptée, notamment *via* le recours à des connaissances de raisonnement dédiées, pour répondre à nos préoccupations.

4.2 Des propositions pour rationaliser la conception des ontologies

Il existe beaucoup de synthèses en matière de construction de ressources ontologiques (par exemple, [UG96, JBV98, Góm99, DSW^{+}99], ou plus récemment [GFC04] qui propose le récapitulatif le plus complet). Avec la profusion d'ateliers qui ont abordé le sujet, on peut bien se rendre compte des incertitudes théoriques et pratiques qui ont marqué la recherche dans ce domaine, lorsque la notion d'ontologie s'est vue adoptée de manière relativement large. Avec le temps, cet engouement semble être retombé, en partie grâce à la meilleurs maîtrise des processus de conception par les acteurs de la discipline, mais aussi faute de consensus. En ce qui concerne

la conception d'ontologies, il n'existe pas en effet de solution unifiée : on a plutôt affaire à un ensemble de solutions locales d'ingénierie, qu'il faut s'efforcer de recenser et de fonder théoriquement, mais dont un utilisateur pourra toujours n'extraire que ce qui concerne ses besoins particuliers.

Dans le cas de la construction des ontologies, à la traditionnelle dichotomie méthodes/outils s'ajoute une grille de lecture relative aux moyens concrets mis en œuvre lors de ce processus. On peut par exemple distinguer :

– l'extraction d'ontologies à partir de textes : ces travaux utilisent des procédés de fouille de textes afin d'en extraire les informations permettant de construire des réseaux sémantiques reflétant des conceptualisations dans la langue. On peut ainsi créer des *proto-ontologies* qui après une phase de complétion et de précision plus ou moins importante de la part des ingénieurs des connaissances, déboucheront sur des ontologies formalisées. Pour un état de l'art et des propositions concrète en la matière, on peut se reporter à [BAC04b] ;
– la fusion et l'adaptation d'ontologies : ici, on vise la reprise de conceptualisations spécifiées dans des ontologies déja existantes liées au domaine que l'on veut conceptualiser. Pour cela, on intègre une ou plusieurs ontologies, placées ou non sur un pied d'égalité, et en en modifiant les spécifications au besoin. Ceci nécessite très souvent une étape d'*alignement*, qui identifie les concepts et les relations que ces ontologies ont en commun. Des exemples de propositions méthodologiques ou techniques concernant cette approche sont trouvables dans ONIONS [GPS99] ou PROMPT [NM03] ;
– le développement collaboratif d'ontologies : ces travaux cherchent davantage à mettre en valeur et assister la nécessaire collaboration entre les concepteurs des ontologies, en mettant à leur disposition des dispositifs de discussion et de gestion de versions différentes des ressources en cours de construction. [Dom98, TNL+05].

Dans le cadre des travaux conduits à l'INA, notamment dans le cadre des réflexions à conduire pour OPALES, notre tâche nous a cependant amené à considérer la manière de développer le contenu des ontologies proprement dit, indépendamment des réflexions sur des moyens auxiliaires. Même si un concepteur peut avoir souvent accès à des ontologies ou d'autres sources concernant des thèmes proches de son activité, on considère ici qu'il ne les réutilise pas directement comme matériau de son travail. Cette hypothèse très artificielle – y compris pour OPALES, puisque Véronique Malaisé, doctorante CIFRE à l'INA, a conduit un travail très important d'extraction de concepts et de relations à partir de textes liés aux applications ! – est utilisée pour se concentrer sur ce qui est censé être le cœur de toute activité de modélisation de connaissances ontologiques : comment l'on fixe une conceptualisation appropriée à une application, et comment on la traduit en un artefact opérationnel.

Il y a dans ce domaine une grande quantité de recommandations quant au contenu formel de l'ontologie. Depuis [BMP+91] jusqu'à [RDH+04], les différents langages de représentation d'ontologies ont fait l'objet de réflexions sur leur usage, sur la manière de créer les meilleures ontologies au moyen des primitives qu'ils offrent. Mais ces études sont trop près des langages implémentés eux-mêmes pour justifier l'ontologie du point de vue d'une application précise et de ses utilisateurs, ce qui comme on l'a vu est l'objectif de toute conception d'ontologie [MSD00]. On a besoin de s'abstraire d'un niveau somme toute proche du niveau symbolique, et d'aller plus vers les connaissances exprimées dans les domaines. De fait, ces deux facettes du développement des ontologies sont évidemment complémentaires, et il va falloir proposer une réflexion méthodologique globale, qui puisse les articuler de manière convenable.

Dans cette perspective, nous nous sommes donc d'abord tourné vers les méthodes génériques de développement d'ontologies, qui visent à rationnaliser le cours de la construction d'ontologies formalisées, de sorte à ce que les concepteurs voient leur tâche assistée par sa décomposition en

plusieurs sous-objectifs plus aisément atteignables.

4.2.1 Des méthodologies pour organiser le cyle de développement des ontologies

De nombreuses propositions de gestion générale du cycle de vie d'une ontologie ont donc pour but d'autoriser la création de spécifications formalisées pertinentes le plus facilement possible. Ces propositions sont souvent inspirées des solutions courantes de génie logiciel, qui s'attachent à la création de programmes à partir de spécifications des fonctionnalités plus ou moins abstraites que les concepteurs de ces programmes doivent mettre en œuvre. Ou bien encore elles prennent pour modèles des méthodologies de conception de SBC génériques – telles CommonKADS [SAA+99] – qu'elles s'efforcent d'adapter au cas spécifique des systèmes ontologiques. Parmi les propositions les plus marquantes, on peut noter celles d'Uschold et Grüninger, celles de l'équipe du LAI de Madrid ainsi que celle du projet On-To-Knowledge.

Nous devons noter que parmi les étapes proposées, nous nous sommes systématiquement intéressé à celles qui *spécifient* le contenu ontologique proprement dit – création des concepts et relations, définition de leur signification naturelle et formelle, encodage des connaissances de raisonnement. C'est en effet de cette phase que dépend tout le reste. L'évaluation, par exemple, est seconde, puisqu'elle prend pour objet le résultat d'un processus de conceptualisation, avant d'en re-déclencher un nouveau si besoin est. Et l'étape de documentation, si elle peut s'effectuer au même moment que la conceptualisation, reste entièrement subordonnée à celle-ci.

Les propositions d'Uschold et Grüninger Cette méthodologie, présentée dans [UG96], regroupe des recommandations introduites dans [UK95] et [GF95], issues de la construction de deux ontologies particulières. De fait, ce travail intègre deux aspects successifs de la construction d'ontologie : un développement informel, suivi d'une conception formelle.

Les étapes principales du processus de développement informel des ontologies sont :

1. la détermination claire du but de la construction de l'ontologie, et du domaine concerné ;
2. la construction de l'ontologie ;
3. l'évaluation du résultat ;
4. la production d'une documentation.

La procédure de construction elle-même comporte une phase de *capture* des connaissances, une phase de *codage* et une phase d'*intégration* d'ontologies existantes. La phase de capture est la plus importante pour essayer de préciser le sens des concepts manipulés : on étudie d'abord la *portée* de l'ontologie, ce qui permet de dégager les *termes* pertinents du domaine, d'effectuer ensuite des *regroupements* des termes qui apparaissent souvent ensemble et de commencer à repérer les termes qui peuvent être rapprochés par similarité ou synonymie. On essaie ensuite de produire des *définitions* des termes utilisés, en utilisant une démarche *middle-out*[5] : il est plus important de définir en premier les termes qui apparaissent comme « cognitivement basiques »

[5]Classiquement, on distingue trois modes de construction des hiérarchies de concepts et de relations :
– *top-down* (descendante) : on commence par introduire les notions dont le niveau de généralité est le plus grand ;
– *bottom-up* (montante) : on commence par les notions les plus spécifiques, et on effectue des regroupements de ces notions pour créer des notions plus générales ;
– *middle-out* : on commence par les notions pertinentes du domaine, que l'on spécialise ou généralise pour obtenir une hiérarchie complète.

dans le domaine[6]. Il faut faire attention à produire des définitions consensuelles les plus complètes possibles et donc éviter les termes ambigus, les définitions circulaires...

Les considérations qui précèdent servent de cadre de compréhension pour un second processus, orienté vers la création d'une ontologie écrite en langage formel, et dont les étapes sont :

1. la précision des *scénarios* dans lesquels va s'inscrire l'utilisation de l'ontologie ;
2. la construction d'une liste de *questions de compétence informelles* auxquelles l'ontologie devra permettre de répondre lors de son utilisation par un système ;
3. la construction d'un ensemble de *termes* – d'abord informels puis formalisés dans un langage formel – utilisés pour formuler des réponses aux questions informelles ;
4. la construction de *questions de compétence formelles* : on formalise les questions informelles en utilisant les termes introduits précédemment ;
5. la spécification des *axiomes*, les lois logiques qui permettent de répondre aux questions formelles ;
6. l'établissement de *théorèmes de complétude* qui définissent un ensemble de conditions nécessaires à l'obtention des solutions complètes aux problèmes posés par les questions de compétence.

La méthode proposée produit donc des ensembles de termes munis d'une structure lâche mais dont les définitions informelles sont supposées guider la compréhension de l'ontologie, et peuvent servir pour la construction d'une ontologie formalisée, résultat le plus élaboré du processus de construction. Les auteurs de l'article précisent que le degré de formalisation à atteindre peut se déduire de l'usage futur de l'ontologie, selon qu'elle soit destinée à être comprise par des êtres humains ou par des systèmes informatiques.

Ce qui nous semble le plus important ici est la détermination des scénarios d'usage et leur association à des questions de compétences informelles puis formelles. Cette stratégie permet de décomposer les objectifs qu'est censée atteindre la conception de l'ontologie : *in fine*, les questions informelles les plus précises sont celles qui permettent d'isoler les éléments du répertoire ontologique (concepts, relations), et les questions formelles permettent de leur attacher une sémantique en logique du premier ordre – ces questions sont des formules logiques utilisant le langage défini par l'ontologie, formules que les axiomes de la dite ontologie doivent impliquer. Malheureusement, ces questions sont évidemment à déterminer pour chaque domaine d'application, et la stratégie d'analyse proposée reste vague – peu d'exemples sont exhibés. De plus, on ne sait pas si une telle approche par question de compétence est propice à ce qui fait le cœur de l'activité de conceptualisation : l'organisation des concepts et des relations les uns par rapport aux autres, en particulier *via* la relation de subsomption. Il n'est absolument pas sûr que rattacher les notions à des questions, fussent-elles organisées selon un schéma objectif/sous-objectif, aide à l'obtention de schémas notion générique/notion spécifique.

Methontology L'équipe du LIA de Madrid a proposé un processus complet de gestion du cycle de vie des ontologies : gestion du projet, développement de l'ontologie, support [FGJ97, BFGG98].

[6]Une telle approche rejoint nos considérations sur l'importance du niveau *structurant* pour une ontologie de domaine (cf. section 3.3.2). [UK95] cite les travaux de Lakoff [Lak87] comme référence à suivre pour la catégorisation : les notions du niveau « basique » sont celles qui sont *primaires* du point de vue de la perception, de l'organisation des connaissances ou encore de l'expression langagière. En particulier, dans le fonctionnement cognitif d'un individu, ces notions sont plus *utilisées* inconsciemment, immédiatement, que *conçues* intellectuellement. L'article reprend l'exemple de la notion basique de *chien*, qui se généralise en *animal* et se spécialise en *retriever*.

Nous ne nous intéressons ici qu'au cœur de la démarche : le développement d'ontologies. Celui-ci est divisé en six tâches :

1. *spécification* : déterminer l'utilisation future de l'ontologie ;
2. *conceptualisation* : obtenir un modèle du domaine au niveau des connaissances ;
3. *formalisation* : transformation du modèle conceptuel en modèle formel ;
4. *intégration* : réutilisation d'autres ontologies ;
5. *implémentation* : construction d'un modèle opératoire utilisable par un ordinateur ;
6. *maintenance* : mise-à-jour de l'ontologie en cas de besoin.

Dans l'idéal, la formalisation et l'implémentation doivent être des étapes de traduction quasi-automatique du modèle qui les précède : le véritable effort de construction a lieu pendant la conceptualisation. Celle-ci repose sur la création de plusieurs structures appelées *représentations intermédiaires*.

Tout d'abord, il faut créer un *glossaire de termes*, que l'on divise en *concepts* et *verbes*. Les concepts vont devoir être regroupés en *arbres de classification* de concepts et les verbes servir à créer des *diagrammes de relations binaires*.

A partir de ces deux structures, on va construire un *dictionnaire des concepts*, qui regroupe toutes les informations concernant lesdits concepts (nom et synonymes, instances, attributs de la classe et de ses instances, relations rattachées au concept). D'autres structures vont également apparaître : table des relations binaires, table des attributs d'instances, table des attributs de classes, table des axiomes logiques, table des constantes, table des formules (pour calculer des valeurs d'attributs), arbres de classification des attributs et table des instances.

Toutes ces structures, destinées à contenir d'abondantes descriptions informelles des objets à représenter dans le SBC, sont supposées cerner au plus près la conceptualisation visée par l'ontologie. Reste qu'on ne sait toujours pas précisément comment obtenir le contenu de ces structures, en particulier les taxonomies de concepts. METHONTOLOGY permet de réaliser un pas appréciable vers plus de rigueur et donc de meilleures ontologies, mais il reste du chemin à parcourir pour guider la conception d'ontologies dont l'organisation serait guidée et justifiée par des principes énoncés spécifiquement pour ce problème.

On-To-Knowledge Dans le cadre du projet européen ON-TO-KNOWLEDGE, l'équipe de l'AIFB de l'université de Karlsruhe a mis au point un cycle de développement et d'évaluation d'ontologies dans le cadre de projets appliqués de gestion de connaissances. Le processus contient cinq étapes génériques : étude de faisabilité, lancement du développement, raffinement, évaluation et maintenance.

Les phases de développement et de raffinement sont celles qui nous intéressent le plus, puisque ce sont celles au cours desquelles les concepts et relations sont élicités, organisés, puis formalisés. Là encore, peu de détails sont donnés sur la manière de définir les notions de l'ontologie et de les organiser les unes par rapport aux autres. Les auteurs ont eux aussi recours aux questions de compétence, mais sans lier ces questions les unes aux autres, comme dans [UG96]. Par contre, ils fournissent plus de détails sur la façon dont les termes issus de ces questions indiquent les candidats-concepts et relations de l'ontologie construite. Si une question comme « Qui est le président de l'entreprise Y » est susceptible d'être posée à un SBC ontologique particulier, on pourra à tout moment exhiber cette question comme justification dans l'ontologie de l'existence des concepts `Président` et `Entreprise`, ainsi que d'une relation d'appartenance valable entre présidents et entreprises.

Cependant, on ne peut pas dire que l'utilisation de telles solutions de conceptualisation soit satisfaisante. En fait, si on suggère au concepteur des pistes pour isoler les notions de l'ontologie ou leur assigner des significations informelles, aucune de ces propositions ne fournit un cadre d'assistance explicite et concret. La phase de conceptualisation proprement dite, celle où les concepts de l'ontologie sont dégagés, définis par certaines propriétés et organisés entre eux, gagnerait à être guidée de manière plus précise qu'elle ne l'est dans ces réflexions. Par exemple, METHONTOLOGY propose un certain nombre de représentations intermédiaires afin de mieux conduire le processus de construction des ontologies au niveau de la connaissance, mais dès que l'on passe au niveau plus fin du contenu de l'ontologie ses préconisations se font plus vagues. Ainsi, ses *arbres de classification de concepts* ne font pas l'objet de beaucoup de recommandations, alors que la création d'une telle hiérarchie est centrale pour l'ensemble de la conception de l'ontologie.

On aurait donc besoin de principes plus précis, permettant d'assurer la cohérence du processus de construction, son adéquation à l'application qui emploiera l'ontologie. Existe-t-il des critères qui garantissent la qualité de la conception, tout en s'insérant dans un cadre méthodologique suffisamment global pour guider suffisamment le concepteur ? [NM01], par exemple, propose un cycle de conceptualisation plutôt complet – et comparable à ceux que nous venons de présenter – mais se pose la question de la correction de la construction de la hiérarchie : est-ce que la hiérarchie est conçue en accord avec l'interprétation courante de la relation de spécialisation, en particulier par rapport à la transitivité de cette relation ? Est-ce que les spécialisations directes d'un même concept sont au même niveau de généralité ? Comment utiliser au mieux les possibilités d'héritage multiple ?

4.2.2 Des principes pour rendre la conception cohérente

Des critères de cohérence isolés Les réponses qui sont apportées à ces questions sont cependant plutôt allusives : elles sont en fait relativement précises d'un point de vue pratique (« les classes d'une même fratrie doivent "être placés sur une même ligne" ») mais ne s'inscrivent pas dans un processus de légitimation explicite et global.

De nombreux articles abordant ce thème ont proposés des principes de cohérence pour légitimer une organisation des connaissances par rapport à ses alternatives. Un même domaine d'application pouvant être modélisé de différentes façons, il était en effet pertinent de fournir des critères permettant d'établir la supériorité d'une modélisation sur une autre.

Gruber, par exemple, propose dans [Gru95] cinq principes de conception à respecter lors de la conceptualisation :

- *clarté* : on veut que les définitions contenues dans l'ontologie ne dépendent pas d'un contexte social ou computationnel particulier – le recours au *formalisme* est encouragé. De plus, une définition complète, c'est-à-dire par condition nécessaire et suffisante, est souhaitable dès que cela est possible ;
- *cohérence* : les raisonnements que l'on peut conduire à l'aide des axiomes d'une ontologie ne doivent pas aboutir à des contradictions ;
- *extensibilité* : l'ontologie doit être conçue de manière à ce qu'une extension consécutive à l'émergence d'un nouvel usage se fasse sans remettre en cause ce qui a été précédemment conçu ;
- *biais d'encodage minimum* : la spécification de l'ontologie doit être aussi indépendante que possible d'un [méta-]langage de représentation particulier ;
- *engagement ontologique minimal* : une ontologie devrait faire le moins d'affirmations possibles concernant le monde modélisé, pour permettre à ceux qui la réutilisent d'en spécialiser les spécifications selon leurs besoins réels.

D'autres approches proposent des principes qui, s'ils s'appliquent à la conception pour des formalismes particuliers, apportent des réflexions intéressantes pour le développement d'ontologies en général. Par exemple, Alan Rector, dans [Rec03], cherche à « normaliser » les ontologies encodées à l'aide de logiques de description. En effet, celles-ci, du fait de la spécification de quantité de définitions exploitables par des classifieurs, présentent souvent une structure hiérarchique complexe, où au final les concepts apparaissent souvent comme spécialisations directes de concepts auxquels ils n'étaient pas rattachés explicitement lors de la modélisation. A l'arrivée, l'utilisateur mis en présence d'une telle ontologie ne peut parvenir à ré-interpréter aisément les motivations conceptuelles qui en ont guidé la création. Rector propose donc de mettre en avant un « squelette » hiérarchique qui regroupe les concepts primitifs[7] et doit répondre à un certain nombre de critères permettant d'atteindre un résultat plus cohérent mais aussi plus accessible : création d'une structure arborescente, homogénéisation des critères de spécialisation à l'intérieur des branches de cette arborescence, distinction entre entités indépendantes – les objets de plein droit du monde physique et conceptuel – et celles qui ne le sont pas – parties ou propriétés des précédentes.

Ces critères sont en fait introduits pour répondre à des besoins communément observés, qui rejoignent en partie ceux introduits par Thomas Gruber. [Rec03] souhaite en particulier :

- minimiser le caractère implicite des distinctions ontologiques : la création de notions se spécialisant les unes les autres obéit à des besoins de différentiation – on retrouve la *differentia* que nous avons évoquée en section 3.2.2. Or les différences considérées relèvent la plupart du temps de considérations qui sont non formelles, et par conséquent ne peuvent se retrouver dans une ontologie encodée *via* une logique de description. Pour limiter ce problème inhérent à l'approche logique, il vaut mieux se borner lors de la conception à n'introduire sous une notion générique des notions spécifiques qui ne relèvent que d'une différentiation unique et aisément identifiable ;
- obtenir une hiérarchie ontologique modulaire et homogène : toutes les différences exprimées dans une sous-branche de l'arbre ontologique doivent pouvoir être rattachées à un même type de différentiation– structurel, fonctionnel.... Ceci permet de faire apparaître au sein d'un même ensemble hiérarchique des notions similaires conceptuellement, ce qui en facilite la compréhension et la maintenance.

Ces critères de conception peuvent parfois se révéler extrêmement utiles. Cependant, outre le fait que nous ne soyons pas entièrement d'accord avec la totalité de ces principes[8] et qu'ils soient parfois flous, il faut remarquer que ces propositions ne suffisent pas à construire un cadre méthodologique cohérent. Comme évoqué dans l'introduction de cette section, on a affaire à des *bonnes pratiques* utiles, mais qui ne constituent pas vraiment un cadre méthodologique général : les éléments de ces solutions n'ont parfois pas de lien clair entre eux. Mais surtout, ces solutions– ou certains de leurs éléments – ne s'appliquent pas à toutes les ontologies, mais à celles qui ont une vocation précise ou qui sont encodées à l'aide d'un langage appartenant à une famille donnée. [Rec03] ne s'en cache pas, qui reconnaît que les méthodes qu'il propose pour la modélisation d'ontologies en LD sont à appliquer sur une hiérarchie qui a déjà fait l'objet d'efforts de rationalisation, tels que ceux que nous allons présenter.

[7] En LD, ce terme désigne les concepts qui sont introduits à l'aide de conditions nécessaires, par opposition aux concepts *définis* qui sont introduits dans des conditions nécessaires et suffisantes.

[8] Ceux de Gruber sont en effet basés sur le fait qu'une ontologie résout d'abord le problème de la modélisation d'un domaine, en s'efforçant de s'abstraire des applications particulières, alors que nous considérons que cet aspect applicatif est au contraire constitutif du besoin de modélisation, et doit donc conduire à l'obtention d'ontologies spécifiant des significations en vue d'applications qui, même si on en considère un large éventail, pré-existent.

OntoClean Les propositions de Nicola Guarino ([GW02], issues d'un très long effort de recherche dont [Gua92], [GCG94], [GW00b], [GW00a] ou [GW01] rendent compte, ont tenté de remédier au problème de la pertinence du contenu des ontologies, en essayant de construire un cadre rigoureux d'analyse formelle qui priviligierait certaines formes d'organisation des hiérarchies ontologiques et en disqualifierait d'autres.

Ces travaux s'intéressent en effet au « nettoyage » des taxonomies, qui sont souvent construites de manière anarchique, et en particulier utilisent abusivement la relation de subsomption. Il s'agit plus ici d'une étape de correction à intégrer dans le processus de développement des ontologies que d'une méthode de construction complète. Néanmoins, les bénéfices de l'application de cette opération peuvent être très grands pour la clarification de ce qui a été spécifié.

L'idée est de s'appuyer sur des *propriétés formelles*, c'est-à-dire des propriétés qui sont supposées refléter les *formes* universelles que la pensée peut attribuer à ses objets. En d'autres termes, il s'agit de recenser un certain nombre de méta-propriétés[9] fondamentales pour notre compréhension des choses. En s'appuyant sur des considérations philosophiques relatives aux modalités d'existence des entités du monde, on construit ainsi une typologie des propriétés (*rôles*, *types*..., cf. [GW00a]) qui peuvent apparaître dans une ontologie, et d'énoncer des règles s'appliquant à la relation de subsomption entre les instances de ces classes de propriétés. C'est la donnée, pour une ontologie en cours de construction, des relations d'instanciation entre propriétés (les concepts de l'ontologie) et méta-propriétés qui va permettre de vérifier que les règles de subsomption ne sont pas violées, et de corriger éventuellement la hiérarchie établie précédemment.

Par exemple, Guarino a recours à la notion de « rigidité » : une propriété est *essentielle* pour une de ses instances si et seulement si elle est applicable à cette instance dans toutes les situations possibles, et une propriété est *rigide* si et seulement si elle est *essentielle* pour toutes ses instances. Une propriété est par contre *anti-rigide* si et seulement si elle n'est nécessaire pour aucune de ses instances : son attribution n'est que possible. Guarino présente les exemples de la propriété `Personne`, qui est rigide, et de la propriété `Agent`, qui est anti-rigide (on peut toujours trouver des situations ou un individu qui a été agent peut ne plus l'être).

La règle de subsomption correspondant à la rigidité stipule qu'une propriété anti-rigide ne peut subsumer qu'une propriété du même type. Par conséquent, on peut améliorer une hiérarchie en attribuant les méta-propriétés de rigidité et d'anti-rigidité aux propriétés concernées et en reconsidérant toutes les relations de subsomption qui enfreignent la règle concernant l'héritage et la rigidité. Au fur et à mesure des corrections, on rendra plus explicites les affirmations sous-entendues par la conceptualisation. Par exemple, `Agent` ne peut généraliser `Personne` : il devra être introduit comme un *rôle* que pourra *jouer* une personne.

En fait, Guarino introduit bien d'autres règles, concernant :
- l'association de conditions d'*identité* – existence d'attributs permettant de distinguer les instances les unes des autres, comme un numéro ISBN pour un livre – aux propriétés,
- la *dépendance* des propriétés entre elles – une propriété est dépendante d'une autre si toutes ses instances sont nécessairement liées à une instance de celle-ci, le concept de `Séquence` dépendant par exemple de celui de `Programme`, toute séquence faisant partie d'un programme – et
- les conditions d'*unité* – permettant de déclarer que tout individu instanciant une propriété peut être considéré comme un *tout* dont on peut reconnaître les parties grâce à un critère

[9]Pour Guarino, les *propriétés* désignent tout ce qui pourra être représenté à l'aide d'un prédicat unaire en logique du premier ordre, ce qui inclut ce que nous avons désigné jusqu'à présent par le vocable de concept. Les *propriétés formelles* sont donc des *méta-propriétés* – énoncées grâce aux logiques temporelle ou modale – qui s'appliquent à ces prédicats unaires.

bien particulier[10].

Ainsi, une propriété dépendante ne peut subsumer une propriété qui ne l'est pas. C'est cependant la rigidité qui joue le rôle le plus important, puisqu'elle permet de reconnaître les *propriétés* fondamentales qui permettent de classer les individus : les *types*.

En résumé, la méthode énoncée dans [GW00a] prend en entrée une taxonomie existante et comporte les étapes suivantes :

1. clarifier les relations d'instanciation de méta-propriétés par les propriétés présentes dans l'ontologie ;
2. vérifier la consistance des ensembles de méta-propriétés rattachés ainsi à chaque propriété : la combinaison obtenue ne doit pas être incohérente ;
3. enlever toutes les propriétés non-rigides : on se concentre ainsi sur les *types* fondamentaux du domaine ;
4. vérifier la validité de chaque relation de subsomption ;
5. ajouter les autres propriétés en vérifiant les contraintes ;
6. regarder s'il manque des concepts dans la hiérarchie.

Le problème est que cette méthode, utilisant des notions relativement compliquées, est assez délicate à mettre en œuvre. Elle l'a pourtant été dans des cas concrets, comme l'analyse de l'ontologie terminologique WORDNET [OGGM02]. Des travaux du laboratoire LARIA d'Amiens – [BKM03], dans la lignée de [Kas99] – ont également appliqué cette méthode pour rationaliser la méthode de modélisation COMMONKADS appliquée au paramétrage d'algorithmes. Et il faut remarquer que l'intégration des principes d'analyse formelle d'OntoClean au sein de l'environnement **WebODE** *via* le module ODECLEAN peut également être utile, en soulageant le concepteur du travail fastidieux de détection d'inconsistance [FG02].

Cependant, comme cette intégration et les travaux effectués le rappellent, cette méthode ne peut s'appliquer que si une hiérarchie a déjà été construite, et que la signification des concepts est suffisamment claire pour attribuer les méta-propriétés de manière certaine. Elle ne donne au fonds pas de réelle *prescription* pour déterminer le contenu des notions ontologiques, tournée qu'elle est vers l'évaluation et la correction formelles. Le premier pas doit toujours être réalisé par le concepteur, seul.

Il faut se rappeler que cette méthode vise plus le niveau de la *forme* absolue des notions que leur contenu sémantique lui-même. Ce contenu sémantique, on l'a vu, est pour les utilisateurs de l'ontologie beaucoup plus lié à la manière dont il est exprimé dans le domaine qu'à sa formalisation. *A fortiori* si cette formalisation s'applique à un niveau plus abstrait encore que celui de la spécification traditionnelle des axiomes logiques traduisant les comportements et propriétés des objets du domaine. Il faut donc toujours trouver un guide méthodologique capable d'assister l'utilisateur pour obtenir une organisation et une signification des primitives de connaissances qui correspondent à l'usage que l'on veut en faire.

Nos expérimentations montrent d'ailleurs que l'emploi de la méthode de Guarino est fortement dépendant de ce stade antérieur de spécification. Nous avons en effet appliqué cette méthode à deux des ontologies réalisées à l'INA pendant notre thèse, celle du cyclisme et de la petite enfance. Ces ontologies étant déjà conçues en accord avec les principes de modélisation rigoureux que nous avons évoqués en section 3.2.2 et dont nous reparlerons par la suite, elles n'ont pas nécessité de corrections significatives, en tout cas au regard des efforts engagés. Mais cela était justement dû à la rigueur de la normalisation sémantique dont elles bénéficiaient.

[10] Les séquences d'un programme ayant par exemple en commun d'avoir fait l'objet d'un choix de montage guidé par une motivation générale qui est celle du programme qui les contient.

En toute généralité, on ne peut disqualifier une méthode qui a démontré son efficacité sur des exemples de conceptualisation *ad hoc* proches de ce à quoi l'on est habituellement confronté dans le domaine. D'autant plus que le niveau de réflexion auquel elle se place est, comme on le verra, complémentaire de notre propre cadre méthodologique.

4.2.3 Initier le processus de conception des ontologies

Des méthodes de vérification formelles peuvent donc fonder l'ontologie d'un point de vue théorique, mais il faut tout d'abord démarrer la conceptualisation avec une organisation des connaissances ontologiques qui soit réellement pertinente pour le domaine.

Pour cela, de nombreuses méthodes se tournent vers l'exploitation des ressources linguistiques que l'on trouve dans le contexte de l'application. En effet, ces ressources, si elles peuvent aider à la compréhension des concepts de l'ontologie (on l'a vu dans le chapitre 3 à la page 3.2.1), sont avant tout utilisées pour déterminer et organiser les concepts et les relations dont on a besoin dans le domaine. Des algorithmes utilisent des connaissances lexicales et syntaxiques pour extraire les termes les plus pertinents des textes relatifs au domaine étudié, termes qui serviront de candidats concepts pour le développement de l'ontologie. Qui plus est, on peut extraire des relations sémantiques entre ces termes – synonymie, hypéronymie[11], méréonymie[12] – ainsi que des relations conceptuelles plus ou moins spécifiques par rapport à l'application concernée : localisation, participation à une action.... Ces relations permettent de placer le terme dans un contexte d'usage structuré ; on a ainsi une ébauche du contexte d'interprétation dont on a vu qu'il était nécessaire pour l'ontologie. De fait, beaucoup de travaux – tels ceux de Véronique Malaisé à l'INA [MZB04] – se sont concentrés sur l'obtention de hiérarchies de termes que l'on peut utiliser comme une première hiérarchie de spécialisation.

Cependant, de même que les termes ne sont pas des concepts[13], les hiérarchies de termes peuvent ne pas correspondre aux hiérarchies de spécialisation de concepts, dont l'interprétation est la subsomption formelle. Si l'on veut obtenir une hiérarchie ontologique à partir des résultats d'extraction de connaissances à partir de textes, il faut valider les propositions issues de l'application des heuristiques de recherche. Il ne peut s'agir que d'un travail semi-automatique, impliquant un part significative de contrôle, de *normalisation* des termes et de la signification qui leur est associée.

On retrouve une situation comparable dans les démarches qui essaient d'initier le processus de développement d'ontologies à l'aide de thesauri. Ces derniers constituent comme on l'a vu une source de légitimité par rapport au domaine, et présentent des similarités structurelles importantes avec les ontologies – du moins les plus simples d'entre elles. Nombreux sont ceux qui ont tenté d'adapter ces structures pour concevoir des ontologies [AFS00, WSWS01, vMS+04, SLL+04]. Cependant, nous avons déjà montré combien diffèrent ontologies et thesauri, notamment du point de vue de la rigueur avec laquelle les liens entre termes sont interprétés. De fait, le passage d'un thesaurus semi-formel à une ontologie formalisée demande des efforts importants. Il faut déterminer puis appliquer des stratégies complexes pour concilier à la fois les liens sémantiques « naturels » du thesaurus et l'interprétation formelle que l'on voudrait obtenir

[11]La relation d'hypéronymie relie un terme spécifique à un terme qui le généralise.

[12]La relation méréonymique (ou méréologique) relie un terme désignant un objet à un terme renvoyant à une partie de cet objet.

[13]Même si ces deux notions sont liées de manière intime. Les contributions de Gilles Kassel [KAB+00] ou de Catherine Roussey [RCP02], tout comme les travaux de recherche effectués à l'INA, en tiennent d'ailleurs compte dans les ontologies qu'ils proposent.

pour ces mêmes liens. Il en va de même pour l'obtention de connaissances de raisonnement ou de liens spécifiques au domaine employables dans des définitions. [SLL+04] montre que cela est envisageable, et même de façon semi-automatique, mais le travail à fournir reste très important.

Une telle adaptation de ressources pré-existantes est donc extrêmement délicate. Tout d'abord, on doit bien faire remarquer que ces ressources ne sont pas toujours immédiatement disponibles : on ne trouvera pas de thesaurus ou de corpus textuel exploitables pour chaque application. De plus, il faut bien faire attention à la façon dont sont adaptés les réseaux de significations que l'on peut récupérer à partir de ces sources. Il faut évidemment fournir un travail de formalisation considérable si l'on veut bénéficier de la rigueur formelle et des connaissances de raisonnement qui constituent les avantages les plus importants des ontologies. Cependant, il ne faut pas négliger l'interprétation naturelle des notions de l'ontologie en construction. Car contrairement à ce que l'on peut espérer, les ressources pré-existantes ne répondent pas forcément aux critères de rigueur que nous souhaitons atteindre, et ce dès le niveau des connaissances « informelles ». Une ontologie a un objectif de classification rationnelle que n'ont pas forcément les péritextes (voir page 20) de l'application qui décrivent très souvent des *faits*, ou les thesauri, qui sont construits en premier lieu pour la recherche d'information. Par conséquent, l'effort de formalisation directe qui a lieu dans la plupart des approches, alors qu'elle prend acte de l'imperfection des ressources dont elle dispose, ne cherche à résoudre ce problème que dans son aspect formel.

La légitimation des notions ontologiques par l'*usage*, permise par l'extraction de ces notions de sources liées à l'application, ne peut faire l'économie du travail de re-construction rigoureuse du contexte d'interprétation nécessaire à la bonne *utilisation* de l'ontologie. Avant de chercher à obtenir des ressources formalisées exploitables par les mécanismes d'inférence d'un SBC, il faut s'assurer de la légitimité ontologique de leur référentiel interprétatif.

4.3 Prescrire une manière de construire les notions ontologiques

Le problème est qu'aucune des contributions méthodologiques que nous venons d'évoquer ne force les concepteurs d'ontologies à expliciter *clairement* le sens qu'ils attribuent aux notions de l'ontologie. En particulier, il y a toujours besoin de fournir un guide méthodologique permettant de légitimer – ou de re-légitimer, si on ré-utilise des hiérarchies existantes – la construction des hiérarchies de spécialisation : si on recommande souvent de *gloser* les notions par le biais d'une documentation ou de commentaires, ce processus reste très informel. Ce qui fait que l'on ne clarifie pas forcément assez son engagement ontologique dans les modalités interprétatives non formelles rencontrées dans le domaine métier.

Une fois encore, nous pouvons utiliser les solutions proposées par Bruno Bachimont. On a vu dans le chapitre 3 (page 88) qu'il était possible de *normaliser* la signification des notions au moyen du langage, de façon à ce que ces notions s'inscrivent dans un réseau interprétable de manière à la fois précise et naturelle. Cette solution peut être reprise dans un cadre méthodologique plus vaste, compatible avec d'autres propositions [BIT02, Bac04a]. Nous rappellerons tout d'abord quels sont les enjeux de cette méthode, de l'élicitation de candidats-concepts à l'obtention d'ontologies opérationnelles. Nous expliquerons ensuite comment le parcours des solutions techniques existantes nous a convaincu pour OPALES d'implémenter cette méthodologie par la réalisation d'un outil adapté, complémentaires des autres environnements de conception ontologiques : **DOE**.

4.3.1 ARCHONTE, un processus de conception d'ontologies régionales

Pour concilier les deux aspects souvent exclusifs de la conception des ontologies – le formel et le naturel – Bruno Bachimont a théorisé leur intégration au sein d'un même cycle de développement qui puisse les articuler de façon cohérente [Bac96, Bac00, Bac01]. Ce cycle comporte trois étapes, illustrées en figure 4.1 :

- *normalisation sémantique*,
- *formalisation*,
- *opérationnalisation.*

Les deux dernières, on va le voir, sont rapprochables de celles proposées dans les autres méthodologies[14], mais la première induit un engagement qui donne à l'ensemble du processus la légitimité métier qui lui manquait[15]. L'exposé qui suit est inspiré de nos travaux présentés dans [Isa01, TI02] ; le rappel méthodologique y était en effet illustré d'exemples et de réflexions complémentaires dégagés des efforts de mise en œuvre et d'application qui ont eu lieu tout au long de notre thèse.

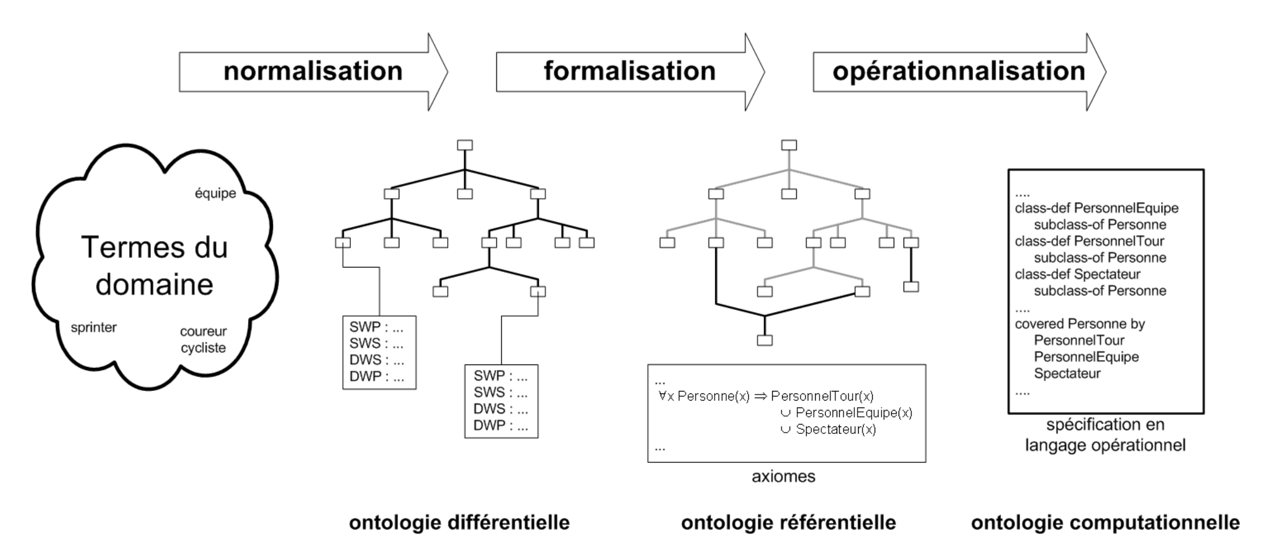


FIG. 4.1 Les étapes de la méthode de construction d'ontologies de Bruno Bachimont, extrait de notre rapport [Isa01]

Normalisation sémantique

L'objectif de cette étape est d'expliciter la signification des libellés linguistiques des concepts pour qu'ils fonctionnent comme des primitives de connaissances, véritables unités de sens compréhensibles et manipulables dans le cadre applicatif visé par le SBC. Bruno Bachimont recommande en effet de se tourner vers la langue pour légitimer, mais aussi servir de point de départ effectif à l'ensemble de la conception de l'ontologie. Pour lui, *« construire une ontologie est un acte de modélisation qui doit s'effectuer à partir de l'expression linguistique des connaissances du domaine ».*

[14] METHONTOLOGY inlut par exemple une *formalisation* et une *implémentation.*

[15] [Bac96] introduit le terme d'ontologies *régionales* pour indiquer le lien entre de telles ontologies et l'application dans laquelle elles sont ainsi ancrées.

Tout d'abord, il faut donc extraire des *termes* des corpus textuels :« textes » écrits ou oraux, documentations ou gloses. Ces termes pourront être utilisés pour désigner les concepts de l'application. Cette étape, antérieure à la phase de normalisation elle-même, n'a pas fait l'objet de recherches dans le cadre de cette thèse, mais de celle de Véronique Malaisé [Mal05].

De tels termes, même si on parvient lors de leur extraction à recréer un réseau interprétatif significatif, ne sont pourtant pas assez précis pour être utilisés comme symboles primitifs du SBC : en l'état actuel des choses, on ne peut être sûr que les procédés d'analyse des textes limiteront suffisamment les interprétations variées dont ces termes peuvent être l'objet. Il faut pour cela qu'un concepteur humain complète le travail d'extraction en fixant une fois pour toute la signification de ces termes, et ce de manière rationnelle : ce que Bachimont appelle l'*engagement sémantique* [Bac00] est à la base de l'*engagement ontologique*.

Pour cela, comme on l'a vu au chapitre 3, dans la section 3.2.2, on a recours à la construction d'une ontologie *différentielle*. On utilise des principes de normalisation linguistique pour expliciter la signification des primitives : similarité avec le père, différence avec le père, similarité avec les frères, différence avec les frères. Et l'on obtient un arbre de notions qui constitue pour chacune d'entre elles un contexte qui lui permet d'être interprétée de manière non ambigue.

Formalisation

Pour concevoir des primitives qui soient exploitables dans un SBC, il faut cependant s'abstraire du paradigme d'interprétation linguistique. Les primitives doivent en effet fonctionner dans des langages de RC dont les mécanismes d'interprétation sont formels ; on devra donc introduire des concepts et des relations munis d'une véritable sémantique formelle. Comme on l'a vu dans le chapitre 2, dans le domaine de la représentation des connaissances ontologiques, très influencé par la logique formelle, c'est la sémantique ensembliste, extensionnelle, qui s'impose. On va ainsi construire une ontologie composée de primitives caractérisées par des prescriptions concernant leur rapport aux objets du monde, leur référence. On parle alors d'*ontologie référentielle*.

Ces primitives formelles n'ont pas intrinséquement de sens interprétatif : elles n'en acquièrent que parce qu'elles sont reliées – par leur libellé – aux notions différentielles. Cependant, ce sont des primitives et elles peuvent grâce aux mécanismes de composition de sens d'une sémantique formelle servir à définir de nouveaux concepts formels. Les concepts étant liés par référence à leur extension, qui est un ensemble d'objets du domaine, on peut avoir recours à des opérations de composition ensembliste (réunion, intersection, complémentaire). On introduit ainsi les axiomes logiques qui régissent les caractérisations des individus qui constituent l'extension des concepts formels. Par exemple, on pourra définir un concept comme étant la disjonction de deux autres concepts, ou bien la négation d'un concept. Dans l'ontologie du cyclisme, les concepts `PersonnelEpreuve`, `PersonnelEquipe` et `Spectateur` forment une partition disjointe[16] du concept `Personne`. On peut aussi exprimer des lois concernant les relations, comme les propriétés algébriques : réflexivité, transitivité... C'est notamment à cette étape que l'on va pouvoir définir les relations par la donnée de leur arité et de leur domaine, ce qui les associera de fait à des produits cartésiens de références de concepts.

La comparaison des extensions permet de définir une relation d'héritage extensionnelle entre les concepts : un concept sera subsumé par un autre si et seulement si son extension est incluse dans celle de son parent. Il faut noter que la hiérarchie de subsomption obtenue peut être différente de celle construite lors de la normalisation sémantique. En effet, les relations d'héritage définies dans l'ontologie différentielle tiennent toujours ici[17], mais l'ajout de nouveaux nœuds,

[16]Leurs extensions sont disjointes et leur union est égale à l'extension du concept qui les subsume.

[17]La relation d'identité entre deux notions est directement transposable du domaine linguistique au domaine

tout comme le passage d'une relation de spécialisation basée sur des similarités et des différences exclusives à une relation interprétée sur une base d'inclusion ensembliste, va modifier la structure arborescente. Si dans l'ontologie référentielle les notions n'étant pas unies par la relation de subsomption s'excluent mutuellement, leurs « transcriptions » en concepts formels peuvent admettre des extensions qui ont un sous-ensemble commun. L'héritage multiple devient donc possible, et l'on peut obtenir une structure de treillis, et non plus d'arbre. Par exemple, on a vu à la page 87 que `Rouleur` et `Grimpeur` étaient deux notions qui étaient en opposition dans notre ontologie différentielle du cyclisme. Cependant, les concepts formels correspondants ont des extensions qui peuvent avoir en commun plusieurs individus (par exemple l'individu `lance_armstrong`). On peut par conséquent définir dans l'ontologie référentielle un concept formel `Grimpeur_Rouleur` dont la référence sera l'intersection des extensions des concepts `Grimpeur` et `Rouleur`, et dont `lance_armstrong` sera un élément. En un sens, le passage à l'ontologie formelle permet de se libérer de certaines contraintes (la manière dont l'interprétation linguistique impose ses règles aux primitives) tout en apportant les siennes (les primitives ont maintenant une signification formelle).

Dans l'ontologie référentielle, on essaie donc de donner aux primitives une signification formelle. Cette signification, si elle doit être en accord avec la signification sémantique naturelle donnée dans l'ontologie différentielle, en est tout de même distincte : l'ontologie référentielle est en quelque sorte le « produit fini » de la conception au niveau des connaissances. Il est clair qu'elle doit expliciter toutes les lois logiques induites par la conceptualisation que l'on veut spécifier.

Opérationnalisation

On dispose donc à présent de tout un système de primitives formelles prêtes à être utilisées dans un SBC. Cependant, les comportements qu'elles prescrivent ne sont compréhensibles que pour les utilisateurs humains, puisque un ordinateur n'a jamais directement accès à la référence des concepts. Ils reste donc à associer à ces primitives des opérations informatiques capables de rendre compte de cette signification référentielle. On appelle *ontologie computationnelle* le résultat de cette spécification opérationnelle.

Les concepts à ce niveau sont définis par les inférences, les calculs que pourra effectuer un système à partir de la donnée des individus qui les instancient dans le monde. Il s'agit de la spécification concrète du fonctionnement du SBC. On utilise pour cela des langages opérationnels de représentation des connaissances[18] qui font appel à des capacités d'inférences précises. Pour un langage de représentation reposant sur les graphes conceptuels, par exemple, ce seront des opérations de manipulation des GC (jointure, calcul de projection, etc. . .). Pour des langages s'inscrivant dans le paradigme des logiques de description, il s'agira plutôt de tests de subsomption entre concepts et de procédures de classification des individus introduits dans la base de connaissances. . . On a vu dans le chapitre 2 (pages 48, 58, 59 et 61) des exemples de spécification dans ces langages opérationnels. Le code 4.1 représente en OWL la propriété de partition disjointe introduite précédemment[19]. On note d'ailleurs à l'occasion que OWL ne présente pas de constructeur permettant de construire une telle disjonction directement : il faut l'exprimer par

extensionnel : elle implique effectivement, pour les deux concepts considérés, que l'extension de l'un est incluse dans celle de l'autre.

[18] Nous en avons évoqué quelques-uns au chapitre 2. [GFC04] en dresse un inventaire complet et récent. Il existe aussi des travaux qui essaient de comparer explicitement leur différentes fonctionnalités, comme [BCKN01].

[19] Cet exemple a l'intérêt d'illustrer l'emploi de la syntaxe XML de OWL, différente de la syntaxe abstraite utilisée dans le chapitre 2.

le biais de plusieurs lois d'exclusion.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<rdf:RDF [...]>
  <owl:Ontology [...]>[...]
  </owl:Ontology>
  [...]
  <owl:Class rdf:about="#Personne">
     <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#PersonnelEpreuve"/>
        <owl:Class rdf:about="#PersonnelEquipe"/>
        <owl:Class rdf:about="#Spectateur"/>
     </owl:unionOf>
  </owl:Class>
  <owl:Class rdf:about="#PersonnelEpreuve">
     <owl:disjointWith rdf:resource="#PersonnelEquipe"/>
     <owl:disjointWith rdf:resource="#Spectateur"/>
   </owl:Class>
  <owl:Class rdf:about="#PersonnelEquipe">
     <owl:disjointWith rdf:resource="#Spectateur"/>
  </owl:Class>
</rdf:RDF>
```

Code 4.1 – Encodage d'une disjonction exclusive en OWL

Il faut remarquer que certains auteurs recommandent d'aller plus loin en ce qui concerne l'opérationnalisation de l'ontologie, en spécifiant le *scénario d'usage* de celle-ci. En l'occurrence, [Für04], on l'a déjà évoqué, distingue quatre types de scénarios :
- inférentiel et implicite : les axiomes sont appliqués automatiquement pour produire de nouvelles connaissances ;
- inférentiel et explicite : les axiomes sont appliqués à la demande de l'utilisateur pour produire de nouvelles connaissances ;
- validation implicite : les axiomes sont appliqués automatiquement pour vérifier la cohérence de la base de connaissances ;
- validation explicite : les axiomes sont appliqués à la demande de l'utilisateur pour vérifier la cohérence de la base de connaissances ;

En pratique, le scénario d'opérationnalisation requis pour les applications que nous visons devrait toujours être le même. Pour que le système contribue à la continuité sémantique, nous cherchons en effet à employer les connaissances de raisonnement à la fois pour contrôler la validité des connaissances de la base et pour produire de nouvelles connaissances permettant de répondre à plus de demandes. De plus, cela doit se faire sans que l'utilisateur ait de charge supplémentaire. Les ontologies devraient donc logiquement être employées dans un scénario d'inférence et de validation implicites.

Cette étape achève donc le processus de spécification d'une conceptualisation. Le passage des éléments d'origine textuelle non contraints à une ontologie différentielle qui prescrit une sémantique interprétative, puis la définition d'une sémantique formelle pour les primitives du langage de représentation, et enfin l'opérationnalisation de ces primitives en vue d'une utilisation

par un SBC concret nous permet de disposer de structures qui permettent de rendre compte d'une conceptualisation de manière claire et cohérente, que ce soit pour les utilisateurs du SBC, pour ses concepteurs ou pour le système lui-même.

4.3.2 Des outils pour faciliter la saisie des spécifications formelles

Après avoir déterminé un cadre méthodologique cohérent, il faut l'instrumenter. Pour le projet OPALES, en effet, l'INA devait produire un environnement de construction d'ontologies compatible avec le format des graphes conceptuels. Nous nous sommes donc tourné vers les solutions existantes, avec à l'esprit les contraintes particulières imposées par la démarche de conception proposée par Bruno Bachimont. Etait-il possible de trouver un éditeur accessible, et compatible avec ces besoins ?

De la même façon que des méthodes de développement génériques ont été proposées pour rationaliser la conception ontologique, celles-ci ont pu s'appuyer sur des environnements élaborés dont l'objectif était de rendre plus intuitive la saisie des connaissances formelles rattachées aux concepts et aux relations. Nous avons déjà partiellement abordé cette problématique dans le chapitre 3, du point de vue de l'assistance à la compréhension des notions formalisées. Nous allons à présent nous intéresser de plus près aux outils effectivement utilisables pour créer des ontologies suivant des approches méthodologiques cohérentes, ou du moins offrant des interfaces intuitives pour créer des connaissances complexes.

Depuis une dizaine d'années, un très grand nombre d'instruments d'assistance à l'édition d'ontologies a été créé. Nous ne nous attacherons pas à tous les examiner en détails ; beaucoup d'études sur ce sujet ont été conduites, de [DSW+99] qui le premier énonçait un jeu de critères dédiés à l'évaluation rationnelle des éditeurs, à [Den04] qui en recense plus de cinquante. Reprenant un travail avancé dans [Isa01] et [TI02], nous nous concentrons ici sur les éditeurs les plus populaires ou marquants. Il faut également mentionner que nous ne nous intéressons ici qu'aux logiciels qui permettent la conception entièrement maîtrisée par l'utilisateur d'ontologies finalisées[20].

Les outils auxquels on peut accéder – une grande partie est disponible gratuitement, du moins en version de démonstration – sont très variés, et diffèrent tant par les interfaces de spécification et de visualisation de concepts et de relations que par la capacité offerte ou non de travailler en ligne, la gestion des évolutions des ontologies créées, l'incorporation à des raisonneurs, ou plus simplement la richesse autorisée pour les spécifications formelles, qui dépend souvent du langage de représentation choisi comme référence pour l'éditeur.

Les plus anciens n'offraient la possibilité d'éditer les ontologies que de manière directe. On spécifiait la plus grande partie des définitions en utilisant un langage de représentation donné, reconnu par l'éditeur. Celui d'entre eux qui a rencontré le plus de succès est **Ontolingua Server** [FFR97]. Ce serveur a été créé par le laboratoire KSL de l'université de Stanford au début des années 90, dans le cadre du DARPA *Knowledge Sharing Effort*, programme pionnier en conception de connaissances partagées. Ontolingua Server cherchait à privilégier le travail collaboratif, plusieurs utilisateurs pouvant se connecter et travailler en même temps sur une ou plusieurs ontologies. Le langage de représentation sous-jacent, KIF, pouvait servir de passerelle entre plusieurs langages implémentés de RC, ce qui permettait un degré de ré-utilisation des connaissances jamais atteint auparavant.

[20]Il existe en effet de nombreuses propositions d'outils permettant d'extraire les concepts et les relations à partir de textes issus du domaine visé par l'ontologie, ou pour créer de nouvelles ontologies à partir d'ontologies déjà existantes. Pour un aperçu de ces plate-formes, on se reportera respectivement à [BAC04b] et à [Kno04].

Mais de nouvelles plate-formes sont apparues, mettant plus l'accent sur la spécification des ontologies *via* des interfaces graphiques adaptées, en minimisant la partie exprimée en langage formel, comme on l'a évoqué dans la section 3.2

Il faut évidemment citer **Protégé**[21] [NFM00], développé par le SMI de Stanford. Protégé est construit autour d'un modèle de connaissances inspiré par le paradigme des frames : *classes*, *slots* (attributs) et *facets* (contraintes sur les attributs) sont les primitives de modélisation proposées. Ce modèle autorise une liberté de conception importante, puisque le contenu des formulaires de spécification des classes peut être modifié suivant les besoins, *via* un système de *méta-classes*, qui constituent des sortes de « patrons » pour les classes du modèle du domaine. L'interface très complète, illustrée dans la figure 3.4 de la page 85, ainsi que l'architecture logicielle ouverte permettant l'insertion de *pluggins*, ont grandement participé au succès de Protégé. En quelques années, cet éditeur a su s'imposer comme la référence, avec une communauté d'utilisateurs extrêmement importante et active. Ses nombreuses extensions lui permettent en particulier de gérer les langages standards comme RDFS et surtout OWL [KFNM04], de créer des axiomes formels de manière intuitive [HNM02], d'accéder aux ontologies par des interfaces graphiques évoluées, etc. Cette prédominance ne pourra qu'être renforcée par le lancement de l'initiative CO-ODE[22], qui a pour objectif la création d'outils d'assistance à la construction d'ontologies OWL riches et cohérentes, et qui se concentre sur cet éditeur.

OILed[23] [BHGS01], a été développé par l'université de Manchester pour éditer des ontologies dans les langages de représentation OIL[24], puis DAML+OIL[25], les précurseurs de OWL. Il est donc explicitement orienté vers la représentation en logique de description expressive, et à ce titre fournit tous les éléments d'interface permettant la spécification des hiérarchies de concepts et de rôles, ainsi que la construction des expressions complexes définissant ces entités. Conçu à l'origine comme un outil simple, il n'a pas d'autre ambition que de donner les moyens de construire des exemples montrant les vertus du langage pour lequel il a été créé. Mais cette simplicité et la robustesse qui en découle, ainsi que le fait qu'il intègre d'office un raisonneur de logique de description, *FaCT*[26], capable de tester la satisfiabilité des ontologies construites ou d'expliciter de nouvelles relations de subsomption entre concepts complexes, en ont fait un outil de référence.

OntoEdit [SEA+02] est un outil propriétaire[27] inspiré par l'approche des frames, mais il gère de nombreux formats libres de la communauté liée au web sémantique (FLogic, DAML+OIL, RDFS), et a le mérite de s'appuyer sur une réflexion méthodologique significative. Il s'est en effet le premier intéressé à la modélisation « naturelle » des axiomes, pour faciliter la traduction d'un langage de représentation à un autre [MSSS00, SEM01]. S'efforçant de mettre en œuvre les propositions du projet On-To-Knowledge, il propose également une gestion originale des questions de compétences, telle qu'on l'a vue en section 4.2.1. Un petit outil d'analyse lexicale compare les termes extraits des différentes questions pour en déduire automatiquement d'éventuelles subsomptions.

[21] Auparavant appelé **Protégé2000**, cet éditeur a repris le nom de l'outil d'acquisition des connaissances qui l'a précédé. Protégé est disponible à `http://protege.stanford.edu/`.

[22] `http://co-ode.man.ac.uk/`.

[23] `http://oiled.man.ac.uk/`.

[24] OIL – *Ontology Inference Layer* – est une proposition du projet On-To-Knowledge. Pour plus d'informations, lire [Ba00]. L'éditeur **OILed** peut être téléchargé à `http://img.cs.man.ac.uk/oil/`.

[25] `http://www.daml.org/language/`.

[26] FAst Classification of Terminologies, `http://www.cs.man.ac.uk/~horrocks`.

[27] Une version de démonstration est disponible sur le site d'Ontoprise, la société qui le développe en collaboration avec l'AIFB de Karlsruhe. `http://www.ontoprise.de`

Le dernier outil évoqué ici sera **WebODE**[28] [ACFG01]. Cette plate-forme en ligne développée par le LAI de Madrid se place au niveau méthodologique dans la lignée d'ODE, un éditeur qui assurait fidèlement le support de METHONTOLOGY, la méthodologie proposée par ce laboratoire. Elle illustre bien l'évolution des outils de construction d'ontologies, puisque les nombreuses tables de *représentations intermédiaires* de son prédécesseur ont été remplacées par une interface plus intuitive, aux dépens des contraintes méthodologiques. L'accent a plus été mis sur la possibilité d'un travail collaboratif ou sur la possibilité, comme dans Protégé, d'étendre la plate-forme par des outils complémentaires, comme un éditeur d'axiomes, un moteur d'inférences, ou bien l'outil **ODEClean**, intégration dans **WebODE** de la méthode de Nicola Guarino (cf. section 4.2.2).

WebODE, similaire en cela aux autres éditeurs, accepte l'export et l'import d'ontologies en RDFS, DAML+OIL et OWL. On observe donc une réelle convergence entre tous ces outils, que ce soit au niveau des modèles de représentation ou de l'encodage des données. L'initiative du web sémantique a conduit à une forte standardisation sur ces points ces dernières années. Néanmoins, certaines fonctionnalités correspondant pourtant à la démarche défendue par les défenseurs de cette initiative ne font pas encore l'unanimité chez les utilisateurs, comme la possiblité d'éditer collaborativement et à distance les ontologies.

Le problème est que cette standardisation ne semble pas aller dans notre sens. Pour mettre en œuvre les propositions méthodologiques retenues pour OPALES, il faut plus que des interfaces de saisie correspondant à des formalismes donnés, fussent-elles extrêmement intuitives et détachées en apparence de ces formalismes. Pour la phase de formalisation cette assistance est utile, puisqu'elle a tendance à rendre la conception indépendante d'un encodage particulier, même si on ne peut effacer le lien privilégié avec le paradigme de représentation qui a inspiré l'outil. L'opérationnalisation est également grandement facilitée, puisque les ontologies sont *exportées* automatiquement dans les langages les plus couramment utilisés. Cependant, on ne voit pas comment mettre en œuvre la phase de normalisation, qui nécessite une gestion relativement fine d'informations semi-formelles, ainsi qu'une co-ordination avec le reste du processus – en particulier la phase de formalisation, comme on le verra après. **Protégé**, avec son mécanisme de méta-classes et de formulaires associés, permettrait bien de faire en sorte que la spécification formelle d'une classe soit accompagnée de la spécification des quatres principes différentiels. Mais il n'est pas sûr que la distinction entre les deux étapes soit suffisamment claire, tout comme il semble être difficile de contrôler la façon dont ces informations sont saisies et répercutées d'une notion à l'autre. Nous avons donc pris la décision de créer notre propre outil, qui prendrait en charge ce que les autres outils – et les méthodes sur lesquelles ils s'appuient – ne font pas.

4.3.3 DOE

DOE – pour *Differential Ontology Editor* – est un outil qui se veut complémentaire des autres éditeurs, orienté prioritairement sur l'étape de normalisation sémantique et de construction d'une ontologie différentielle. Il a tout d'abord été réalisé pour les besoins d'OPALES, qui demandait comme on l'a déjà dit un outil pour créer des ontologies qui soient tout à la fois compréhensibles par des utilisateurs non experts en RC et servir de *supports* pour la création de graphes conceptuels, le formalisme de RC retenu. Nous avons développé **DOE** en collaboration avec Raphaël Troncy au tout début de notre thèse, et nombre des réflexions qui suivent sont tirées de nos premières publications [TI02, BIT02].

[28] http ://delicias.dia.fi.upm.es/webODE/

Création d'une ontologie différentielle dans DOE

La mise en œuvre de l'étape de normalisation est ce qui a motivé en grande partie le développement de cet éditeur. Il attache donc un soin tout particulier à la modélisation des notions différentielles. De fait, même si ce n'est qu'un outil de saisie – les principes différentiels sont exprimés en langue – nous avons essayé d'en faire un outil de saisie *rationalisé*.

Tout d'abord, nous avons recensé les différentes fonctionnalités permettant la création d'une ontologie différentielle. Pour répondre à ces besoins, nous avons créé une interface (cf. figure 4.2) comprenant une vision arborescente des notions – une pour chaque type de notion, concepts ou relations – ainsi qu'un formulaire de saisie présentant dans des champs textuels les quatre principes différentiels, anglicisés :

– *Similarity with parent* – SWP ;
– *Similarity with siblings* – SWS ;
– *Difference with siblings* – DWS ;
– *Difference with parent* – DWP.

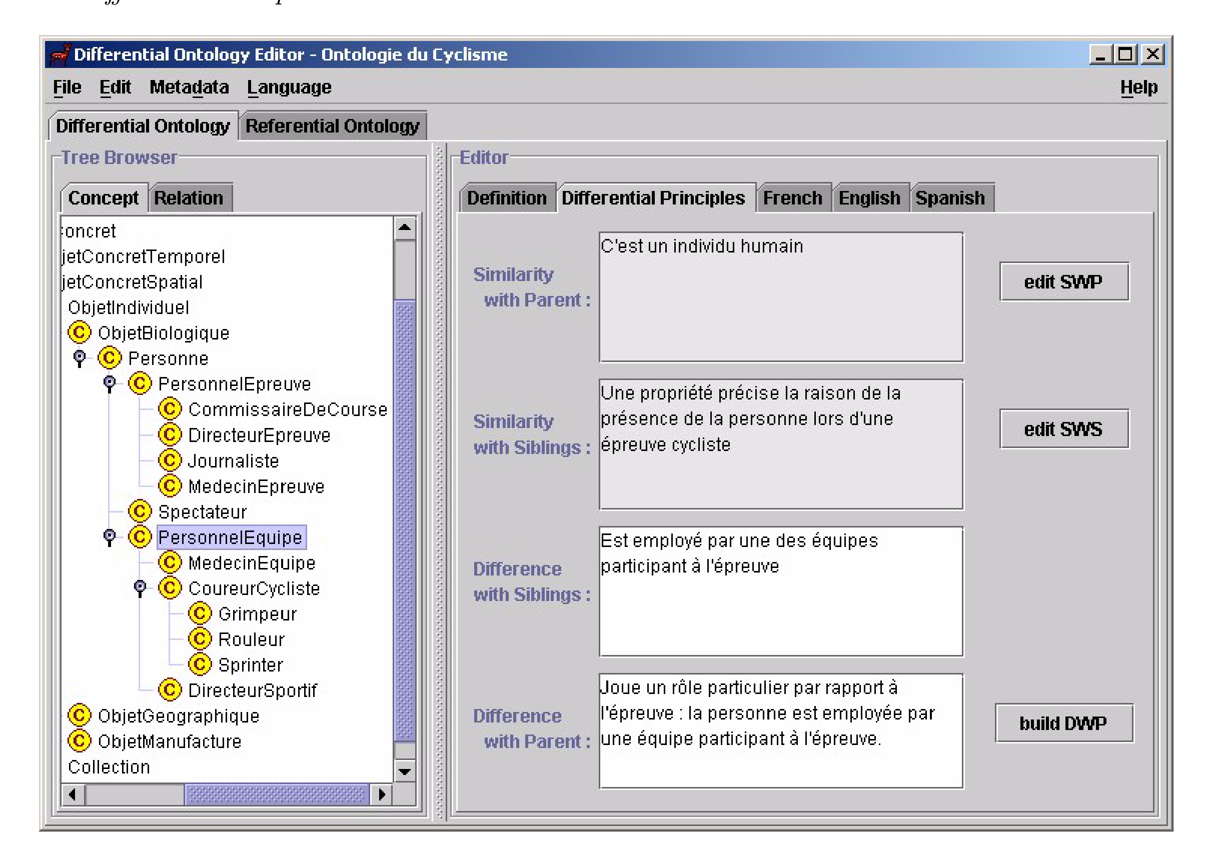


Fig. 4.2 Les principes différentiels de la notion `PersonnelEquipe` dans *DOE*

Il faut mentionner qu'outre les possibilités de création et de modification de notions différentielles, notre éditeur permet une gestion terminologique minimale : à chaque notion, et pour plusieurs langues[29], on peut associer un *terme préféré*, et l'on peut spécifier des synonymes de ce terme. Il est également possible de préciser une *définition encyclopédique*. On complète ainsi

[29]Cette gestion de plusieurs langues, pour certaines des informations attachées aux notions, n'a pas pour ambition de rendre l'éditeur réellement « multilingue ». Ce problème est en effet bien plus complexe ; des notions verbalisables dans une langue pourraient ne pas l'être immédiatement dans une autre. L'éditeur se borne donc

l'ancrage textuel de l'ontologie différentielle, en autorisant un lien explicite entre la notion et ses manifestations langagières. C'est aussi un moyen d'articuler l'étape de normalisation avec les éventuels résultats d'outils d'extraction de termes...

En ce qui concerne la création du contenu conceptuel des notions différentielles, l'outil développé se propose d'assister l'utilisateur dans cette saisie, voire d'en automatiser une partie. En effet, au cours de nos expériences de conception d'ontologies différentielles, quelques faits intéressants ont pu être observés.

En premier lieu, il y a souvent partage de SWP et SWS par l'ensemble des concepts membres d'une même fratrie. En effet, toutes les notions d'une fratrie sont filles d'un même père, et toutes admettent une propriété discriminante déterminée à partir d'un axe sémantique commun : cela est nécessaire à la cohérence de la structure locale de l'ontologie. Par exemple, dans le domaine de la petite enfance, le travail de Véronique Malaisé a conduit à la distinction, parmi les différentes actions à considérer pour l'application, entre des actions entreprises au nom d'un enjeu collectif, comme des rites d'intégration ou de rejet, et des actions personnelles, parmi lesquelles on compte les différents types de soins prodigués à un nouveau-né. Les spécialisations introduites immédiatement sous `Action` sont donc `ActionPrescriteSocialement` et `ActionsPersonnelles`. Celles-ci ont bien toutes les deux en commun avec leur notion parente le fait de définir des actions particulières et concrètes[30] : leur SWP seront identiques. Il en va de même pour les valeurs de leur attribut SWS : on ne va parvenir à distinguer un type d'action de l'autre que parce qu'on les considère *tous les deux* du point de vue de l'axe sémantique que constitue leur *motivation*. Cette motivation, à ce niveau de la conceptualisation différentielle, ne pourra être que « personnelle » – on définit alors le DWS de `ActionPersonnelle` par rapport à `ActionPrescriteSocialement` – ou « sociale/culturelle » – on définit alors le DWS de `ActionPrescriteSocialement` par rapport à `ActionPersonnelle`.

L'éditeur remplit donc automatiquement les deux champs SWP et SWS pour toute une fratrie dès lors qu'une des notions possède l'information. Si l'ontologiste veut déplacer une notion dans la taxinomie, les deux principes de similarité de celle-ci sont adaptés à sa nouvelle fratrie. Une telle décision d'implémentation peut paraître contraignante. Mais elle découle des principes mêmes de la méthode de spécification adoptée, qu'elle aide ainsi à appliquer correctement.

Ensuite, le principe de différence avec le père (DWP) apparaît souvent comme la somme de la spécification de l'axe commun aux frères (SWS) et de celle de la différence avec les frères (DWS). En effet, on se donne d'abord le moyen de créer une différence, que l'on concrétise ensuite pour achever la définition du concept. Ainsi, pour les actions dont nous venons de parler, la différence avec le concept englobant d'`Action` provient bien du fait que l'on s'est concentré sur l'axe de la motivation, et que l'on ait effectivement situé chacune des spécialisation sur cet axe – *via* les valeurs spécifiques « motivée socialement » et « motivée par une cause personnelle ». L'éditeur propose donc de construire lui-même la différence avec le père (DWP) en concaténant simplement les énoncés des principes DWS et SWS. Mais ici, il ne doit pas y avoir de contrainte définitive : l'utilisateur a tout loisir de changer ce qui ne reste qu'une simple suggestion de l'outil.

Enfin, il faut signaler que les principes sont souvent considérés par l'utilisateur dans le même ordre, à savoir SWP, SWS, DWS et DWP. On commence effectivement par préciser que la notion éditée est bien subsumée par une autre. On affirme ensuite ce qui fonde son originalité : choisir le fils plutôt que le père, c'est insister sur une caractéristique donnée (SWS) et en indiquer une

actuellement à donner la possibilité de mettre en correspondance des termes et des définitions informelles. De plus, il faut rappeler qu'un certain nombre d'informations centrales (les principes différentiels, notamment) ne peuvent être explicitées que dans une seule langue, qui prendra alors le statut de langue de référence pour l'ontologie. Toutes les langues ne sont donc pas traitées de la même manière.

[30] Par opposition à des `Pratiques`, qui ont une vocation plus globale.

valeur possible (DWS) qui l'oppose aux autres frères, et donc qui différencie ses instances de toutes les autres instances du père. L'interface de l'éditeur reflète ce cheminement, comme le montre la figure 4.2, qui illustre l'exemple déroulé jusqu'à présent (notamment aux pages 89 et 125) pour le concept `PersonnelEquipe`.

On a vu que la modélisation différentielle d'une notion permettait de définir celle-ci par l'interprétation de sa position dans la structure hiérarchique globale de l'ontologie. A titre informatif, l'éditeur génère donc, dans l'onglet *Definition*, une *définition différentielle* des notions, en affichant le chemin menant de la racine de l'arbre à la notion considérée et en récapitulant les valeurs des différences avec les pères rencontrées sur ce chemin.

Création d'une ontologie référentielle dans DOE

Gestion des informations référentielles L'objectif de la formalisation est d'attacher aux notions une signification formelle permettant de fixer la référence de ces notions, qui pourra être traduite ensuite dans un langage compréhensible par un SBC.

La première opération de cette étape consiste à récupérer les deux taxinomies établies pour l'ontologie différentielle. Aucune des informations apportées dans l'ontologie référentielle ne peut en effet remettre en cause la hiérarchie de subsomption : les deux arbres de concepts et de relations construits précédemment réapparaissent donc tels quels dans *DOE*.

L'utilisateur peut ensuite ajouter des concepts ou des relations qui ne présenteront plus les informations sémantiques des notions différentielles. L'héritage multiple étant possible, l'éditeur tient à jour, pour chaque entité, la liste de ses parents et permet d'en ajouter ou d'en retirer.

Nous avons choisi dès le début de différencier ces relations des concepts correspondant aux entités du domaine. Cela a une répercussion au niveau de l'ontologie référentielle. La description de la référence des relations est effectivement spécifique : pour une relation donnée, il faut pouvoir préciser son arité ainsi que ses domaines. L'éditeur propose de rentrer ces informations, et le choix des domaines se fait parmi la liste des concepts déjà entrés dans l'ontologie. **DOE** prend également en charge des vérifications basiques, en fonction des relations de spécialisation entre relations : l'arité, quand elle a été définie, doit être propagée correctement le long de la hiérarchie de subsomption. De même, dans **DOE**, les domaines d'une relation spécialisent les domaines de la relation qu'elle spécialise. Ainsi, en cas d'héritage multiple, les domaines de la relation fille doivent être des sous-concepts des domaines de toutes les relations parentes.

Il faut également mentionner qu'en matière de cohérence des spécifications, notre éditeur prend en compte les relations d'inter-définition entre les entités référentielles lors des modifications de l'ontologie. Ainsi, il sera impossible de supprimer un concept qui est employé comme domaine d'une relation. De même, si une notion référentielle est supprimée, toutes les notions que cette suppression rend orphelines sont également retirées, sauf si par exemple cela vient à contredire la règle précisée auparavant.

Articulation des étapes différentielle et référentielle La méthodologie que nous avons retenue insiste bien sur les différences entre les deux premières étapes. Les objets qui y sont manipulés sont de nature différente (linguistique pour les uns, logique pour les autres), ce qui a des répercussions sur la manière dont on les construit. Il ne faut cependant pas oublier qu'il y a des liens forts entre ontologie différentielle et ontologie référentielle : les notions différentielles donnent une signification linguistique aux concepts référentiels auxquels ils sont directement associés. La conséquence directe est d'ailleurs la reprise dans l'ontologie référentielle de la hiérarchie de subsomption établie dans l'ontologie différentielle. Une question s'est alors posée : l'éditeur doit-il présenter ces deux étapes de front, ou l'une après l'autre ?

D'après notre expérience, il semble préférable de construire l'ontologie référentielle au fur et à mesure de l'élaboration de la taxinomie différentielle. En effet, les notions qui y sont présentes ont pour but de préciser le sens des entités référentielles correspondantes. Mais ces entités doivent pouvoir servir à l'introduction d'autres entités (concepts ou relations) dès qu'elles sont définies. Attendre que toute la taxinomie différentielle soit entièrement terminée pour introduire des concepts référentiels qui auraient pu l'être bien plus tôt peut se révéler préjudiciable à la sérénité supposée du concepteur d'ontologies. Par exemple, le concept `Grimpeur_Rouleur` de la page 126 devrait pouvoir être défini dès que les concept `Grimpeur` et `Grimpeur` sont introduits *via* l'ontologie différentielle. L'éditeur permet donc de mener ces deux étapes simultanément en passant à tout moment d'une ontologie en construction à l'autre.

Il faut alors souligner les problèmes de cohérence que cela induit. Si une évolution monotone de l'ontologie différentielle ne pose pas de problèmes (les relations de subsomption restent valides dans l'ontologie référentielle), le déplacement ou la suppression d'une de ses branches peut créer des concepts référentiels « orphelins » n'étant plus subsumés par un quelconque concept. L'éditeur prend en charge cette vérification, et informe l'utilisateur le cas échéant. Enfin, il est à noter que l'introduction des entités référentielles ne permet pas de remettre en cause les données de l'ontologie différentielle puisque ces dernières sont issues d'une étape qui est logiquement antérieure.

De l'articulation entre ontologies différentielle et référentielle, on passe logiquement à l'articulation entre ontologies réferentielle et computationnelle. De fait, celle-ci requiert un niveau de détail formel dont on a eu un aperçu dans le chapitre 2 et qui n'est pas pris en compte par notre éditeur. Nous comptons pour cela sur des fonctionnalités d'*export* qui si elle peuvent permettre d'obtenir des ontologies computationnelles élémentaires permettent surtout de poursuivre le travail de formalisation dans des environnements de conception plus complets sur ce point.

Opérationnalisation, exports et re-formalisation

Comme on l'a vu dans les paragraphes qui précèdent, les entités référentielles introduites dans *DOE* restent en effet *primitives*. La possibilité de définir des concepts par conditions nécessaires et suffisantes avec des combinaisons booléennes, ou encore par l'expression de contraintes relationnelles, n'est en effet pas proposée. D'autres outils (comme **Protégé** ou **OilEd**) pouvant assurer le support de telles fonctionnalités, nous préférons recourir à des mécanismes d'exports adaptés plutôt que de refaire ce qui est très bien pris en compte ailleurs. Il en va de même pour les axiomes qui s'appliquent aux relations qu'ils soient simples (symétrie, réflexivité. . .) ou plus complexes (règles de composition). Pour cela, on peut utiliser un outil comme **TooCom**, développé par Frédéric Fürst [Für04] : Cet outil, en plus de permettre l'édition intuitive des règles exprimables dans le formalisme des GC, permet les différents types d'opérationnalisation que nous avons présentés en page 127. Ce genre de co-opération entre outils d'édition s'inscrit tout à fait dans l'optique du web Sémantique[31], chacun s'efforçant de réaliser correctement ce qu'il fait de mieux, et de le communiquer du mieux possible aux outils les plus pertinents pour ce qu'il ne sait pas faire.

Une telle opportunité aurait d'ailleurs pu arriver dans le cadre du projet OPALES : lorsque le projet a été lancé, il était convenu que notre éditeur permettrait l'édition d'ontologies dont le niveau d'expressivité correspondrait à celui des supports classiques de GC. En l'occurrence, ces supports, comme on l'a vu dans le chapitre 2, comprennent la spécification des types de

[31]Voir par exemple l'ensemble des expérimentations réalisées dans le cadre du millésime 2003 de l'atelier *EON* (*Evaluation of Ontology Tools*), `http://CEUR-WS.org/Vol-87/`.

concepts et de relations, des liens de subsomption entre ces types, des domaines et co-domaines des relations, ainsi que des marqueurs individuels instanciant les types de concepts. Toutes ces possibilités font partie des fonctionnalités de **DOE**, comme nous l'avons vu. Et un export des spécifications de l'ontologie formelle en CGXML (comme illustré par le code 2.1 de la page 48) permettait d'obtenir facilement les ontologies computationnelles recherchées. Cependant, il est apparu très rapidement que pour une exploitation correcte des index conceptuels il fallait tenir compte de connaissances de raisonnement qui correspondaient à une extension du modèle minimal des GC : les règles de raisonnement [Sal97]. Il aurait donc fallu créer une interface de saisie de telles règles. L'existence d'un outil comme **TooCom** nous délivre d'un tel fardeau, puisqu'il permet d'importer notre ontologie en CGXML, et de la compléter à l'aide des règles pertinentes pour nos applications[32].

DOE gère l'export de ses ontologies référentielles dans les (méta-)langages de représentation les plus couramment rencontrés dans l'initiative du web sémantique : RDF(S) [RDF04b], OIL [FvH+01], DAML+OIL[33] et OWL [OWL04]. En fait, **DOE** enregistre les ontologies différentielles et référentielles dans son propre format, qui utilise la syntaxe de XML [XML04]. Et, comme nous le reverrons dans le chapitre suivant, il applique ensuite une feuille de style XSLT [XSL99], ce qui permet d'opérer une traduction syntaxique du format de **DOE** à celui dans lequel on veut que l'ontologie soit encodée.

Les ontologies référentielles exportées dans les formats retenus sont donc *importables* dans la majorité des outils adaptés à la formalisation et à l'opérationnalisation, notamment ceux que nous avons présentés en section 4.3.2. Nous avons donc rendu notre éditeur interopérable avec les outils les plus fréquemment rencontrés. Car il faut également mentionner que ce qui est valable pour l'export l'est aussi pour l'import : on peut traduire des ontologies d'un format standard à celui qui est utilisé pour **DOE**. Ainsi, on pourra récupérer des ontologies déjà existantes – ou des proto-ontologies (thesauri ou réseaux de termes) – dans **DOE** et conduire le travail de normalisation sémantique qui leur est nécessaire, commencer la formalisation et éventuellement de la conclure dans un outil plus approprié.

Le problème est que cette forme d'interopérabilité est nécessairement limitée par l'expressivité des différents outils et langages mis en œuvre. Par exemple, l'expressivité formelle de OWL est considérablement plus importante que celle autorisée par **DOE**. De fait, toutes les possibilités formelles de notre éditeur – spécification de la relation de spécialisation, définition des domaines et co-domaines des relations – sont incluses dans celles de OWL, alors que la plupart des possibilités – définitions complexes ensemblistes, axiomes relationnels simples – de ce langage issu des logiques de description ne sont pas prises en compte dans notre modèle. Cela n'est en principe pas gênant si on se contente d'exporter des ontologies construites avec **DOE** vers d'autres éditeurs, mais on peut perdre quantité d'informations si on importe dans **DOE** des ontologies formellement détaillées.

Plus grave encore est la possibilité de perdre les informations définies pour l'ontologie différentielle : les langages de représentation d'ontologies ne sont pas forcément conçus pour exprimer les informations linguistiques que l'on assigne aux notions de cette ontologie. Or ces informations doivent être exportées, si on veut conserver les significations normalisées dans l'ontologie différentielle, et même si on ne peut le faire que pour les concepts et relations introduits dès l'étape de normalisation. Il a donc fallu trouver des solutions de remplacement, telles que l'inscription des principes différentiels dans les commentaires – habituellement non structurés – des langages

[32]Notre honnêteté nous poussera tout de même à reconnaître qu'au moment de la création des ontologies dans Opales, l'outil de Frédéric Fürst n'était pas encore disponible. Et que la spécification des règles de raisonnement a donc dû se faire à l'aide d'une interface autrement plus frustre...

[33]`http://www.daml.org`.

courants. Les expérimentations que nous détaillerons au chapitre 5 montrent qu'il est tout à fait possible de parvenir à des compromis acceptables en la matière.

Nous avons donc vu dans cette section comment il était possible d'utiliser une méthodologie proposant véritablement une assistance à la création et l'organisation du contenu conceptuel des notions ontologiques, tel que ce contenu est conçu et compris dans le domaine lui-même. La méthodologie proposée par Bruno Bachimont offre en effet un cadre complet permettant de concilier la définition de primitives de représentation de connaissances au niveau naturel et leur formalisation dans un cadre logique. Nous nous sommes donc tourné vers cette proposition, que nous avons essayé de mettre en œuvre.

Notre contribution a un double aspect méthodologique et technique. Nous nous sommes efforcé, à travers son application à des ontologies concrètes, de déterminer quelles étaient les formes d'assistance que l'on pouvait produire pour faciliter le travail de normalisation sémantique, et comment on pouvait articuler concrètement les ontologies différentielles, référentielles et computationnelles. Ces efforts ont débouché sur la création d'un outil de conception, **DOE**, qui n'est certes qu'un outil d'édition, mais qui a bénéficié d'un effort certain de rationalisation en vue de faciliter l'appréhension de la méthodologie retenue. Cet outil offre les fonctionnalités basiques qu'on est en droit d'attendre pour effectuer la normalisation et la prolonger par une formalisation et une opération élémentaires. Il ne se place en effet pas ouvertement en concurrence avec les autres outils, dont on a vu que les points forts concernaient justement ces deux dernières étapes. Au contraire, il essaie d'attendre un niveau d'interopérabilité satisfaisant, permettant de bénéficier des avancées réalisées sur l'ensemble du processus de développement d'ontologies.

De fait, notre outil semble avoir intéressé nombre de personnes. Accessible librement sur le site d'Opales[34] dès sa création, il a été téléchargé près de six cents fois en trois ans. A défaut de savoir s'il a comblé d'aise une partie significative de ce public, nous pouvons d'ores et déjà affirmer qu'il constitue une réponse concrète à des questions largement partagées au sein de la communauté de l'ingénierie des connaissances.

Nous avons donc un cadre de conception cohérent pour les ontologies, qui forment le vocabulaire référentiel de compréhension de la substance de nos index. Nous avons cependant vu qu'en ce qui concerne la forme des index, c'est-à-dire l'utilisation effective des concepts et des relations au sein de structures décrivant le contenu des documents, il était souhaitable d'adjoindre aux ontologies des *patrons d'indexation*. Or les réflexions que nous avons présentées jusqu'ici ne s'en sont pas préoccupées. Alors qu'il ne semble pas illégitime de se poser la question du lien entre la conception de ces patrons et celle du contenu conceptuel et formel des ontologies auxquelles ils sont tant liés...

4.4 Ingénierie ontologique et patrons d'indexation

Les patrons d'indexation sont des patrons d'*utilisation effective* de l'ontologie : ils formalisent la *situation* des notions de l'ontologie dans un usage applicatif. De fait, ils constituent des réponses aux questions de compétence que l'on a vues en section 4.2.1, puisqu'ils indiquent les descriptions que les ontologies doivent autoriser dans le cadre du SBC. Nous allons voir comment on peut les introduire dès l'étape de conception de ces ontologies, en amont de leur utilisation dans le SBC. On va en particulier montrer comment on peut lier ces patrons à des *patrons de conception*, de sorte que la conception bénéficie à la fois d'une légitimité théorique liée à l'adhésion à des

[34] `http://opales.ina.fr/public`

principes propices à consensus et d'une reconnaissance « métier » apportée par une prise en charge correcte des structures de description typiques.

La notion de patron de conception, importée depuis la communauté du génie logiciel, est en effet une idée de plus en plus populaire dans la communauté de l'ingénierie ontologique. Cette notion a cependant été adaptée de différentes manières. Après avoir brièvement abordé quelques points de vue constradictoires sur le sujet, nous détaillerons la proposition d'Aldo Gangemi et de ses collègues : celle-ci introduit, dans le cadre d'ontologies de haut niveau, des patrons de conception qui sont spécialisés en vue d'applications concrètes. Nous montrerons que cette approche présente cependant des imperfections, que nous tenterons de corriger par une prise en compte mieux appropriée des patrons d'indexation.

4.4.1 Ingénierie ontologique et patrons de conception

Les *Design Patterns* [GHJV95] du génie logiciel ont pour ambition de présenter des solutions attestées à des problèmes de conception logicielle courants dans le domaine de la programmation orientée objet. Servant de guides pratiques pour les développeurs, ils décrivent des conceptions logicielles consensuelles à un niveau plus abstrait que celui des classes et des instances normalement utilisées dans les systèmes implémentés. Un tel effort d'*abstraction* est en effet nécessaire au *consensus* recherché, et donc autorise une plus grande *ré-utilisabilité*.

Ces objectifs généraux sont de fait relativement proches de ce que l'on recherche en représentation de connaissances ontologiques : des spécifications de conceptualisation partagées et réutilisables, mais aussi exploitables dans des systèmes concrets. Inspirés par ces similarités, des chercheurs ont donc proposé d'adapter ces concepts de génie logiciel à l'ingénierie ontologique.

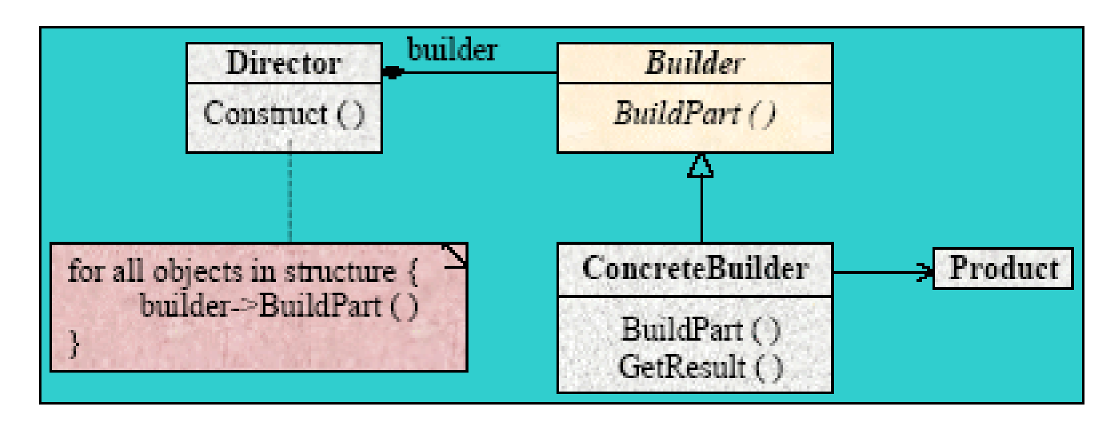


FIG. 4.3 Structure du patron *Constructeur* employé dans [Dev99]

Les premières contributions, reconnaissant la pertinence du contenu de patrons de conception logiciels particuliers, ont directement essayé de les utiliser comme matériau pour leurs ontologies. [Dev99, GFP02]. Par exemple, [Dev99] utilise un patron *Constructeur*, qui permet de représenter un processus de construction logicielle de manière distincte de cette construction elle-même, dans le cadre de génération d'explications de ce processus. Ce patron (voir figure 4.3) introduit les classes de *directeur* (la classe qui lance la construction), *constructeur* (une interface générique auquel accède le directeur), *constructeur concret* (celui qui prend effectivement en charge la construction) et *produit*. Ces classes, ainsi que les liens qui les unissent – héritage ou utilisation – sont utilisées de façon directe dans une ontologie dédiée à la communication homme-machine, qui comportera un `Client`, un `Générateur` et un `Générateur concret`. Dans ces approches, les

patrons sont au même niveau que les ontologies qui s'en inspirent. Si [Dev99] réutilise le patron *Constructeur*, c'est qu'il veut réaliser une ontologie sur la *génération* logicielle. Dans ce cas, les patrons existants sont considérés comme des proto-ontologies : ils contiennent des connaissances qui initient le processus de spécification de conceptualisations, mais ils n'en sont qu'une amorce qui doit faire l'objet d'un travail de précision et de formalisation approprié.

D'autres travaux ont essayé de développer leurs propres *patrons de conception ontologiques*, cherchant à reproduire l'approche des patrons de conception logiciels sans nécessairement utiliser des design patterns existants comme point de départ.

Un premier courant s'est attaché à la description d'architectures représentationnelles pour les ontologies. Ce niveau est donc plus abstrait que celui que l'on vient de voir, puisqu'on se préoccupe de l'organisation des connaissances ontologiques et non de leur contenu proprement dit. Reprenant la terminologie de [Gua95] (se référant lui-même à [Bra79]), nous employons pour les désigner le vocable de patron de conception ontologiques *épistémiques*.

[Rei99], ainsi, propose une série de patrons en vue de la représentation et de la gestion d'ontologies terminologiques, afin d'améliorer la réutilisabilité d'ontologies médicales qui doivent permettre la conception de définitions ou d'expressions terminologiques complexes. Pour cela, des patrons sont proposés, qui fournissent par exemple les outils nécessaires à la composition de *chaînes d'expression* tenant à jour leur *contexte* (en l'occurrence, une information à représenter) : une *expression* (qui se décompose en *expressions concrètes*) maintient un lien vers ce contexte, tout en permettant d'accéder à son successeur dans la liste des autres expressions qui peuvent exprimer le contexte. Le contexte « virus infectant une bactérie » peut ainsi être désigné par les expressions « phage » puis « bactériophage ». Comme nous l'avons fait remarquer, on a plus à faire ici à des structures facilitant la modélisation (on est encore à un niveau relativement informel) de l'ontologie, en vue d'une application donnée, que d'agencements des éléments spécifiques du contenu de celle-ci.

C'est une vision similaire, quoique plus tournée vers la formalisation, qui guide les travaux cherchant à déterminer des patrons pour la conception des axiomes ontologiques. [SEM01] propose par exemple des *patrons sémantiques* pour représenter des connaissances de raisonnement indépendamment de langages particuliers. On peut ainsi faciliter le partage et la réutilisation d'axiomes dont les conséquences sont semblables alors que leur encodage diffère. De fait, ces patrons sont définis par les contraintes qu'ils imposent aux connaissances assertionnelles de la BC, que ce soit en *entrée*, pour qu'ils puissent s'appliquer, ou en *sortie*, puisque leur application produit de nouvelles assertions, ou bien des contraintes sur les assertions existantes. [SEM01] fournit l'exemple du patron *range-local inverse*, ou une relation est définie comme l'inverse d'une seconde, mais uniquement lorsque celle-ci prend ses valeurs dans un co-domaine particulier[35]. Mais d'autres patrons sont évoqués, comme celui du raisonnement partie-tout, de l'héritage avec exception.... On a donc affaire à des schémas d'axiomes exprimés au niveau de la connaissance, mais traduits en termes de besoins concrets vis-à-vis de la gestion d'une BC. Des traductions opérationnelles de ces axiomes sont ensuite à fournir, suivant le méta-langage de représentation retenu, et bien sûr en fonction des concepts et relations de l'ontologie à qui ils s'appliquent.

[HNM02] propose une approche plus proche des langages formels classiques. Les patrons que ces auteurs présentent sont en fait des énoncés, au niveau de la connaissance, de schémas d'axiomes qui sont de fait assez semblable à la forme logique de ceux-ci, et non donnés sous

[35]Par exemple, pour le point de vue d'analyse audiovisuelle d'OPALES, la relation `aPourTraitSaillant` a pour co-domaines des objets audiovisuels « indépendants » (`Séquence`, `Image`) ou bien des propriétés « dépendantes » (caractéristique visuelles ou sonores, résultat d'une action de montage ou de filmage). Son inverse local pourra être la relation `estContenuDans` dans le premiers cas, et la relation `estUneCaractéristiqueDe` dans le second.

la forme de contraintes sur le contenu des BC. Au cours d'une étude d'un corpus d'ontologies formalisées, les auteurs de cet article ont recueilli les formes les plus couramment utilisées pour ces axiomes. On a ainsi des « patrons » comme « Toutes les instances de la classe {C} ne contiennent pas la valeur {V} dans le slot {S} ». Ces patrons inspirent ensuite la création de formulaires destinés à faciliter leur saisie dans l'outil **Protégé**, et la traduction de cette saisie dans un langage de représentation.

Ces méthodes constituent des assistances utiles à la capture de connaissances de raisonnement. Cependant, leur position au niveau *épistémologique* en fait des structures assez éloignées de la problématique de l'usage concret d'une ontologie. Il n'y a pas vraiment de prescription d'un contenu axiomatique particulier, juste des indications quant aux *formes* que celui-ci peut prendre. Par conséquent, on ne peut pas dire qu'instancier de tels schémas constitue un acte d'engagement ontologique fort, propice à une quelconque légitimation des connaissances axiomatiques qui sont produites lors de cette instanciation.

Lorsqu'on utilise un *design pattern*, on se place pourtant dans un cadre précis, celui de composants architecturaux qui ont une signification en termes métier, fût-elle mentionnée de manière extrêmement abstraite. Si nous sommes intéressé à une adaptation de cette vision dans l'ingénierie ontologique, c'est que ce que nous recherchons, ce sont des éléments qui relèvent plus du contenu des connaissances ontologiques que de leur structure[36].

Plus proches de notre questionnement sont donc les méthodes qui essaient de définir des patrons conceptuels génériques (nous les appellerons donc patrons de conception ontologiques *conceptuels*), éventuellement munis d'axiomes, que l'on peut directement spécialiser pour obtenir les connaissances ontologiques au niveau d'un domaine ou d'une application. Ces travaux essaient d'initier la conception avec des composants issus d'ontologies de haut niveau satisfaisant des critères :

- d'utilisabilité : les connaissances génériques doivent être facilement mobilisables lors de la production des ontologies, et leur contenu doit être adapté à un éventail applicatif considéré à présent comme prioritaire par rapport aux problèmes formels – tels que discutés dans la section 4.2.2 – ou épistémiques ;
- de vraisemblance : les notions de haut niveau utilisées reposent sur des travaux de conceptualisation de référence. Ces travaux peuvent s'appuyer eux-mêmes sur des principes philosophiques éprouvés, ou être adaptées de théories abstraites largement adoptées, comme les théories mathématiques.

On compte ainsi rendre plus explicite l'engagement ontologique plaçant la conceptualisation de l'application dans une vision « propre » du monde qui l'environne.

[CTP00], par exemple, présente la manière dont on peut reprendre des théories abstraites pour construire des ontologies dédiées à des domaines particuliers. Ces théories introduisent en effet des concepts, des relations et des axiomes dont on peut faire *hériter* – cette approche se place dans le cadre des *frames* – des concepts et des relations plus spécifiques, mais aussi des axiomes qui réutilisent la structure et le contenu – en les spécialisant – des axiomes génériques. Ainsi, la théorie des graphes acycliques introduit des *nœuds*, des relations correspondant à l'existence d'un *arc* ou à l'*accessibilité* d'un nœud depuis un autre, et des axiomes applicables à ces entités, tel que celui qui permet le calcul de la relation d'accessibilité à partir des arcs. Cette théorie peut être spécialisée en une théorie des graphes bloquables, ou les nœuds peuvent être bloquants

[36]Il faut justement ici rappeler la différence entre la structure de l'expression des connaissances ontologiques et la structure du contenu conceptuelle des descriptions. La forme des axiomes ontologiques n'est évidemment pas l'agencement des concepts et des relations dans les assertions effectivement produites au moyen de l'ontologie…

vis-à-vis de la relation d'accessibilité. A son tour, cette théorie sera spécialisée en une théorie des réseaux de distribution, où les nœuds sont à présent vus comme *producteurs*, *consommateurs* ou *intermédiaires*, et où des flux d'*éléments transportables* peuvent être fournis aux consommateurs si ceux-ci sont accessibles. . . . Finalement, on obtiendra ainsi une théorie, celles des circuits électriques, qui va fournir les concepts et relations de haut niveau pour toute ontologie abordant ce domaine, munis d'une axiomatisation adaptée.

[BLS$^+$00] introduit des mécanismes semblables, dans le cadre d'ontologies géographiques *multi-couches*. Les ontologies susceptibles d'être utiles dans ce domaine nécessitent en effet des connaissances issues de domaines spécialisés – on parle de couches – multiples : réseaux, couverture spatiale. . . . Pour faciliter la conception de tels artefacts, Benslimane et ses collègues proposent de distinguer, à l'intérieur de chaque couche, le niveau *fonctionnel* du niveau du *domaine*. Le niveau fonctionnel regroupe les connaissances abstraites qui fondent les théories auxquelles sont rattachées chacune des couches. On peut ainsi retrouver une théorie de la couverture d'un espace qui va introduire des concepts et des relations fonctionnels (zones, relations entre ces zones), ainsi qu'un certain nombre de contraintes (partition, complétude ou incomplétude de la couverture d'un espace, continuité) et d'opérations (calcul de voisinage, agrégation de zones) qui peuvent être activés pour cette théorie. De fait, lors du passage du niveau fonctionnel à celui du domaine (la représentation du cadastre, ou le découpage zones de couverture hospitalières), on va non seulement spécialiser les notions de la théorie fonctionnelle (les parcelles sont des zones) mais aussi choisir de sélectionner ou non les opérations et les contraintes, en fonction du domaine. Par exemple, dans le domaine du découpage hospitalier, les zones devront réaliser une couverture totale de l'espace, et on devra connaître pour chaque zone l'ensemble de ses voisines. Les niveaux fonctionnels contiennent une quantité très restreinte de notions, mais celles-ci sont liées d'un niveau à l'autre : les zones d'une théorie de la couverture spatiale sont d'un certain point de vue assimilables aux nœuds d'un réseau, ce qui permettra par exemple d'appliquer à la relation d'accessibilité entre zone des contraintes issues des propriétés de l'accessibilité dans la théorie des réseaux.

On semble assez loin, dans ces approches, de trouver des patrons d'indexation que nous avons introduits, qui sont très liés à la notion de *motif* relationnel récurrent. Mais force est de reconnaître que le fait que les patrons sémantiques dont il est question dans ces travaux soient introduits comme des *théories* est loin d'être contradictoire avec l'acception des *design patterns* dans le génie logiciel, qui insiste bien sur les méthodes applicables aux objets présentés[37]. Et qu'on a bien là, à un niveau abstrait, les prémisses de ce que seront les connaissances assertionnelles d'une BC. Le typage fonctionnel de ces entités indique bien les liens et opérations possibles, et, de fait, souhaitables, dans le cadre applicatif que l'on a retenu.

Le rapprochement avec nos patrons d'utilisation devient plus explicite dans la présentation que font Gangemi et ses collègues des patrons de conceptions ontologiques appliqués à la création d'ontologies *noyaux* [GM03, MGOS04, GCB04].

Gangemi, dans [GM03], propose en effet de réutiliser les notions de haut niveau de l'ontologie DOLCE [GGM$^+$02], obéissant aux principes de cohérence formelle énoncés par Guarino, au sein d'une structure relationnelle qui présente ces notions dans un contexte d'utilisation – ici, les « Descriptions et Situations ». Le patron qui en résulte (*D&S*, cf. figure 4.4) présente des *descriptions* de *séquences d'événements* qui ordonnent des entités temporelles (*perdurants*), des *rôles* que des entités physiques (*endurants*) peuvent jouer dans ces événements, ainsi que des

[37]Et nos patrons d'indexation, après tout, sont bien eux aussi accompagnés de connaissances de raisonnement, même si *a priori* leur légitimation vient de l'application qui les exploite.

paramètres qui sont utilisés pour décrire rôles et événements, et prennent leurs valeurs dans des *régions* plus ou moins abstraites. Cette structure générique cherche à s'abstraire d'un domaine particulier tout en résolvant un problème de description ontologique donné. En effet, ces patrons de conception ont bien vocation à rendre compte de « bonnes pratiques »[38] de modélisation, censées être adaptées à des usages existants.

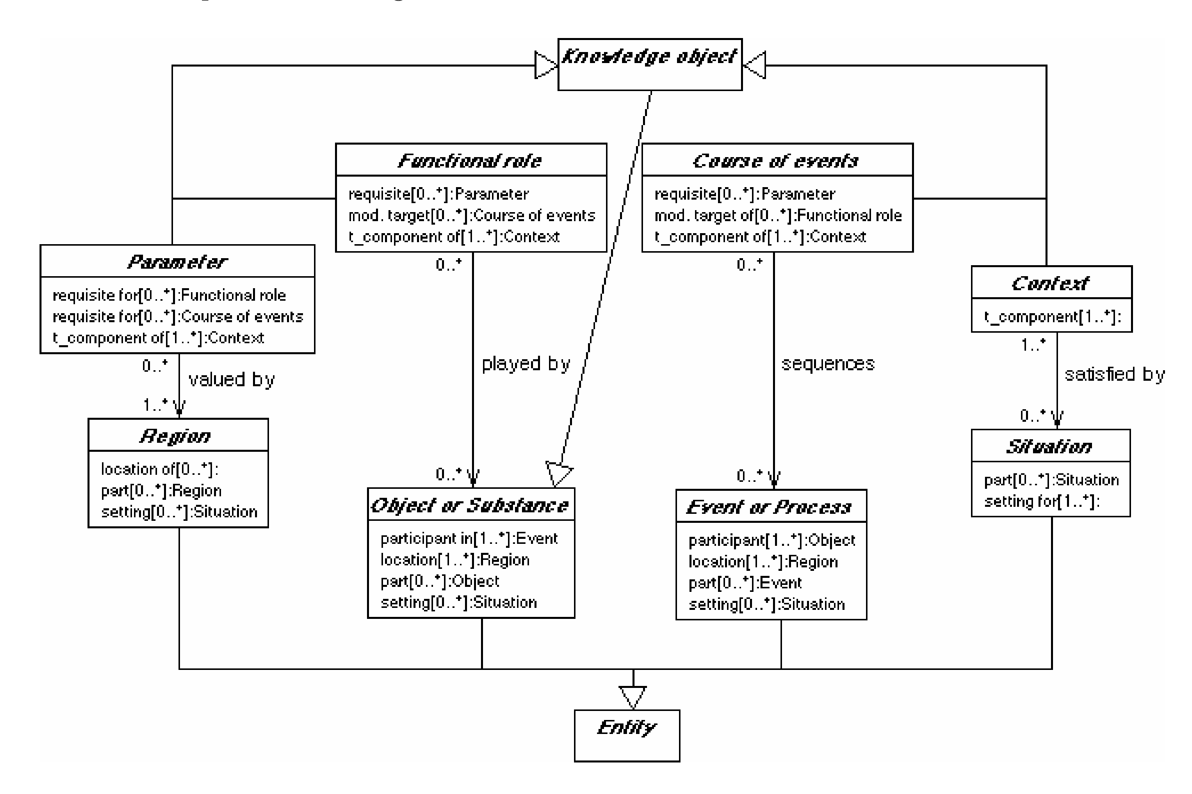


FIG. 4.4 Patron de conception « Descriptions & Situations », extrait de [GM03]

Il faut également remarquer que si les patrons viennent bien avec des axiomes logiques – ceux qui sont rattachés à ses concepts et relations dans DOLCE – ils demeurent indépendants de langages de RC particuliers. Les auteurs mentionnent également parmi les critères que doivent valider un patron sa capacité à être représentable de façon intuitive et compacte. On augmente ainsi la capacité du patron à être aisément compris et repris.

Une telle structure est de fait aisément adaptable à la conceptualisation d'un domaine particulier, pourvu que l'usage visé par cette nouvelle conceptualisation soit similaire à celui qui a dicté l'élaboration du patron. Il suffit alors de rattacher aux notions qu'elle introduit celles, plus spécifiques, du domaine d'application envisagé. Cela se fait en spécialisant – par subsomption classique – les concepts présents dans le patron de conception. La figure 4.5 montre comment les concepts de D&S ont été ainsi spécialisés pour s'adapter au domaine de la description de processus d'inflammation et de leur résultats [GCB04] : on transpose tout d'abord les concepts génériques du patron en des concepts qui ont toujours une valeur fonctionnelle, mais qui relève du domaine modélisé, avant d'introduire les concepts plus concrets qui seront réellement employés dans les descriptions.

[38]Un groupe de travail du W3C relatif aux « bonnes pratiques du web sémantiques » les considère d'ailleurs comme l'une de ses pistes de recherche principales. Voir `http://www.w3.org/2001/sw/BestPractices/.` pour plus d'informations.

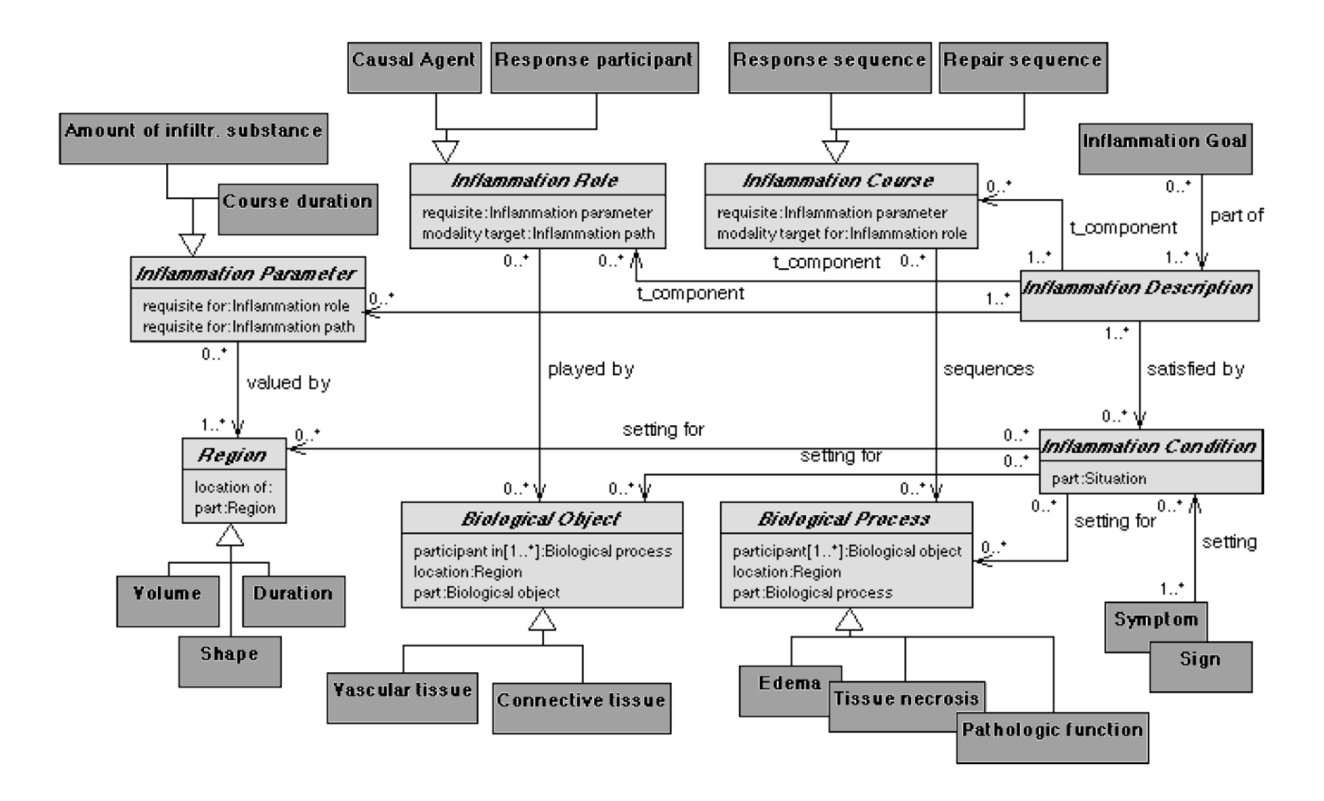


FIG. 4.5 Spécialisation du patron D&S pour le domaine de l'inflammation, extrait de [GCB04]

L'emploi de concepts grisés marque la différence entre les concepts généraux du domaine et le niveau « métier » plus concret

Dans cette approche, le niveau auquel sont spécialisés les concepts reste assez abstrait, même si les concepts relèvent clairement du domaine. L'objectif est en effet d'obtenir une modélisation réduite aux notions les plus générales du domaine, dite *noyau*[39]. Cette ontologie peut être considérée comme le plus grand dénominateur conceptuel commun à toutes les ontologies qui concerneront le domaine d'un point de vue plus appliqué. La figure 4.5 montre en effet comment la transposition du patron de conception dans le domaine biomédical des inflammations – les notions du centre de la figure – va être à son tour spécialisée pour apporter des notions proches des besoins des applications – les notions de la périphérie de la figure. De fait, la spécialisation consistera à remédier au second plan l'aspect « fonctionnel » des objets du patron spécialisé, pour mettre en avant des objets « métier » pertinents.

Rattacher une ontologie de domaine à un patron de conception de haut niveau améliore indéniablement sa qualité intrinsèque : on clarifie l'engagement ontologique en adhérant explicitement à une conceptualisation qui a bénéficié d'une effort théorique important, et qui de plus apporte des connaissances de raisonnement se répercutant, *via* la relation de subsomption, aux notions du domaine. D&S, par exemple, par l'intermédiaire des notions de l'ontologie DOLCE, emprunte à un travail de conceptualisation et de formalisation qui s'appuie lui-même sur nombre d'analyses philosophiques, linguistiques et cognitives, comme détaillé dans [MBG+02]. Une telle approche facilite de fait la réutilisation d'éléments de conceptualisation particuliers (vocabulaire, agencement de concepts et de relations), y compris celle de connaissances de raisonnement, pour assister le développement d'ontologies. Les structures proposées sont en effet également supposées avoir été conçues en fonction d'usages attestés, et donc être facilement adaptables à des cas particuliers.

Cependant, on ne peut se prononcer sur la manière dont les usages concrets dans les SBC sont pris en compte dans de telles approches : l'engagement par rapport aux ressources fondamentales des ontologies de haut niveau est effectivement très important, puisque la méthode privilégie un processus de spécialisation « brute » des patrons pendant la conception. Les usages sont bien pris en compte, mais lors de la conception du patron lui-même. Et au moins dans le cas de D&S, on retrouve une structure très proche de ce qui est présent dans l'ontologie de haut niveau – DOLCE. On se place donc dans ce qui ressemble à une approche descendante *top-down* classique. Que se passe-t-il si des besoins applicatifs particuliers s'écartent de ceux qui ont été anticipés pour le patron ?

4.4.2 Patrons de conception de haut niveau et besoins applicatifs

Nous avons tenté de reproduire cette démarche de spécialisation d'un patron de conception ontologique pour une ontologie dédiée à la description des documents audiovisuels [IT04]. Cette ontologie, dont le développement a répondu aux besoins dégagés lors de la thèse de Raphaël Troncy, mais aussi, dans une certaine mesure, à ceux d'un des points de vue du projet OPALES[40], se concentre sur la description des documents dans une perspective que l'on essaie de rendre la plus neutre possible. En l'occurrence, l'indexation effectuée à l'INA nous a semblé un bon point de départ, puisque cette indexation essaie de rendre compte de la forme et des dispositifs audiovisuels en tentant d'anticiper un maximum de besoins.

[39]Pour un récapitulatif des enjeux de la conception des ontologies noyaux, on peut se reporter à [VB96]. L'atelier de la conférence EKAW 2004 dédié à ce type d'ontologies, auquel nous avons participé, donne lui un bon aperçu des recherches actuelles s'inspirant de cette approche [GB04].

[40]Le point de vue de lecture critique des documentaires géographiques avait en effet besoin d'outils de catégorisation des éléments audiovisuels.

L'obtention d'une ontologie noyau a été guidée à la fois par ces besoins et par les notions de haut niveau introduites par Gangemi et ses collègues. Le patron D&S, conçu pour un usage générique de description, semblait en effet convenir à nos besoins. Pour adapter ces notions au domaine audiovisuel, il a fallu appliquer[41] le patron aux deux activités structurant le cycle de vie du document audiovisuel de télévision, avant son archivage : la production et la diffusion. Ainsi, du point de vue de la diffusion, la description d'un document audiovisuel suppose un enchaînement d'événements de diffusion, tels que l'*émission* d'un programme sur un *canal de diffusion*. Des rôles de diffusion, comme *diffuseur* ou *récepteur*, sont joués par des entités qui peuvent être des *organisations* ou des *personnes*. Et, dans la description de ces événements, nous trouvons des paramètres comme la *date de diffusion* ou le *taux d'audience*, valués respectivement par des *dates* et des *taux*. La figure 4.6, que nous avons adaptée de notre article méthodologique [IBL05], illustre la manière dont ces concepts spécialisent ceux du patron de conception pour obtenir un patron au niveau de l'ontologie noyau.

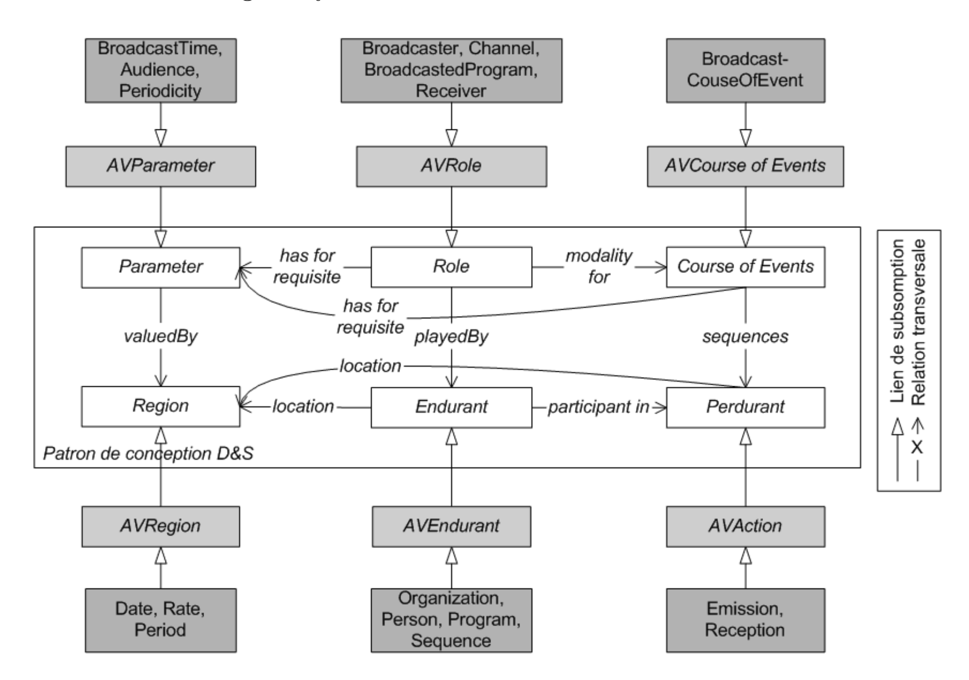


FIG. 4.6 Introduction de concepts généraux de l'audiovisuel suivant le patron de conception D&S

La vue du patron D&S a été simplifiée pour la lisibilité de l'ensemble

Pour autant, une telle structure sera-t-elle utilisable par l'indexeur ou l'expert du domaine ? La complexité des notions impliquées par des considérations abstraites peut masquer la vue applicative sur le domaine, et limiter la pertinence de l'ontologie.

Par exemple, notre étude des besoins en matière de description documentaire audiovisuelle (reposant en grande partie sur l'examen des notices produites à l'INA et concernant à ce titre autant le catalogage que l'indexation) aboutit à l'obtention d'un patron d'utilisation dont la structure relationnelle, illustrée en figure 4.7, est beaucoup plus simple que celle esquissée en figure 4.6 [IT04]. Les descriptions dont nous avons réellement besoin sont centrées plus sur les documents que sur les situations de production ou de diffusion : certaines subtilités informa-

[41] La patron D&S place en effet au cœur de la conceptualisation qu'il propose la notion de séquence d'événements temporels.

tionnelles du patron de conception ne seront donc pas nécessairement à expliciter pour notre application. Par exemple, savoir qu'un document donné joue le rôle de programme diffusé dans une séquence d'événements de diffusion est relativement inutile pour nous. On préférera plutôt des propriétés simples permettant de relier directement le programme à sa date de diffusion, sa mesure d'audience, etc. Le patron spécialisé ne correspond donc pas à ce qui est effectivement demandé pour l'indexation. Qui plus est, cette lacune est surtout visible au niveau *relationnel*, puisque d'une part les chemins relationnels qui sont proposés sont plus longs que ceux dont on a besoin, et, d'autre part, le niveau lui-même auquel sont données ces relations est très abstrait. On risque donc de complexifier la tâche des futurs utilisateurs, et ce sur le point même qui rendait les solutions ontologiques intéressantes par rapport aux solutions d'indexation classiques.

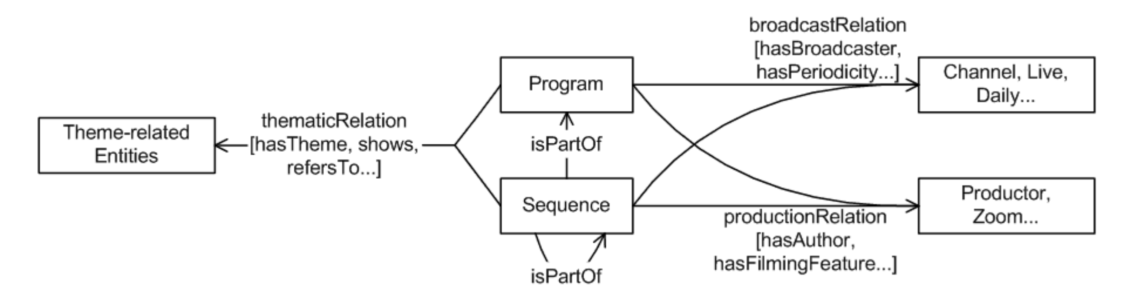


Fig. 4.7 Patron de description d'un document audiovisuel

On peut en fait se demander si les considérations qui guident les patrons ontologiques, en se concentrant sur la cohérence au service de la réutilisabilité, ne sont pas biaisées par l'importance trop grande qu'elles accordent aux principes théoriques rigoureux des ontologies fondamentales. Comme [Men03] le souligne, les approches concernant les ontologies de référence sont souvent inspirées par une démarche *réaliste* – concentrée sur la détermination de modèles les plus proches d'une réalité supposée – et ignorent souvent les considérations *pragmatiques*, qui ancrent une ontologie dans une tâche donnée : on recherche une modélisation de la réalité, sans se poser la question de l'utilisabilité d'une telle modélisation – ou même du bien-fondé d'une telle quête. De fait, certaines approches essaient d'« adoucir » leur engagement métaphysique en incluant dans leur préoccupations des cas de description conceptuelle génériques. C'est notamment le cas de DOLCE et du patron D&S, qui s'efforcent d'obtenir des éléments de modélisation inspirés par les usages ainsi que des considérations cognitives ou linguistiques supposées plus proches de ceux-ci. Mais comme on vient de le voir, si ceci peut suffire à obtenir une ontologie de domaine plus facilement adaptable à différentes applications, on n'est pas sûr de pouvoir rendre compte de tous les aménagements pragmatiques qui font d'une ontologie applicative la ressource adaptée au fonctionnement concret d'un SBC particulier.

La légitimité d'une ontologie est en effet également liée à sa facilité d'utilisation, et à sa capacité à permettre la conception de connaissances assertionnelles qui permettent de répondre efficacement aux questions de compétence. De fait, l'application stricte de principes méthodologiques tels que ceux que nous venons de voir aide à répondre à des questions de compétence, mais du type de celles qui s'intéressent plus à l'ontologie en tant que telle qu'au comportement du système dans lequel celle-ci sera insérée : est-elle cohérente ? Est-ce qu'elle rend compte convenablement des lois du monde ? Est-ce qu'on peut la partager facilement ?

Se contenter de spécialiser un patron de conception ne suffit donc pas toujours à obtenir de véritables *patrons d'utilisation*, tels que les motifs d'indexation que nous avons évoqués tout au long de ce manuscrit. On doit faire la part entre les décisions théoriques d'un patron de conception

quant à l'organisation du monde à décrire, et le domaine, qui peut exiger une modélisation plus pragmatique. Dans un cadre appliqué concret, il faut veiller à se placer au niveau des notions typiques de l'application – qui ne sont pas nécessairement les notions générales du domaine – et chercher à obtenir une structure vraiment proche du besoin descriptif.

C'est pourquoi nous pensons que le paramétrage des conceptualisations tel que présenté dans [CTP00, BLS+00], permettant de sélectionner ou non les éléments de modélisation introduits dans les théories génériques, est quelque chose à considérer de très près pour la conception des ontologies. L'approche descendante doit être nuancée par la considération explicite des patrons d'utilisation réels de l'application, et des connaissances qui leur sont attachées.

4.4.3 Vers une solution articulant patrons de conception et patrons d'utilisation

Nous ambitionnons donc d'inclure les patrons d'utilisation ontologiques dans le processus d'adaptation des patrons de conception ontologiques (que nous pourrons appeler par la suite patrons de conception). De fait, la spécialisation de ces patrons de conception produit des motifs qui sont utilisables dans le domaine, mais qui peuvent rester trop abstraits pour une application donnée, qui emploie évidemment ses propres concepts spécialisés, mais qui peut aussi recourir à un agencement relationnel différent de celui-ci. Il faut donc faire la distinction entre le patron issu de la spécialisation directe du patron de conception – spécialisation que nous appellerons patron d'utilisation ontologique noyau, ou patron noyau – et celui qui sera utilisé pour l'application – désigné par la suite par le vocable de *patron d'utilisation ontologique applicatif*, ou patron applicatif. Concevoir un cadre d'application des patrons de conception qui soit pertinent pour une application revient alors à articuler convenablement le patron noyau avec le patron applicatif. Pour bénéficier, *via* le patron noyau, de la légitimité apportée par le patron de conception de haut niveau, il faut faire en sorte que le patron applicatif soit rattaché par spécialisation logique au dit patron noyau, qu'il en « dérive » logiquement comme le patron noyau dérive du patron de conception (cf. figure 4.8). Mais cela ne doit absolument pas remettre en cause les structures et les notions du patron applicatif, garantes de la pertinence de l'ensemble du processus par rapport à l'utilisation finale de l'ontologie.

Spécialisation des notions

Le patron d'utilisation applicatif doit spécialiser le patron d'utilisation noyau, cela est nécessaire tant pour clarifier l'engagement ontologique qui caractérise le PUA que pour bénéficier des connaissances formelles apportées par le patron de conception.

On sait quel est le niveau du patron d'utilisation applicatif : ce qui est valable pour le patron d'indexation, qui est un cas particulier de patron applicatif, est généralisable. Au chapitre 3, en effet, on a vu que les patrons d'indexation étaient donnés au niveau *structurant* pour l'application. Cela reste évidemment valable pour des cas qui ne relèveraient pas forcément de l'indexation : une application va toujours avec des notions qui structurent de manière privilégiée la conception que l'on peut s'en faire et les pratiques que l'on y rencontre. On y trouvera donc des notions conceptuelles ou relationnelles d'inspiration *métier*.

On connaît également le niveau auquel sont donnés les concepts et les relations du patron de conception. Celui-ci vise en effet la conceptualisation d'un usage relativement précis, mais indépendamment d'un domaine concret. On trouvera donc des notions issues d'une ontologie de haut niveau, marquées par une connotation fonctionnelle – qui renvoie aux raisons de la présence des notions dans le patron qui les a sélectionnées.

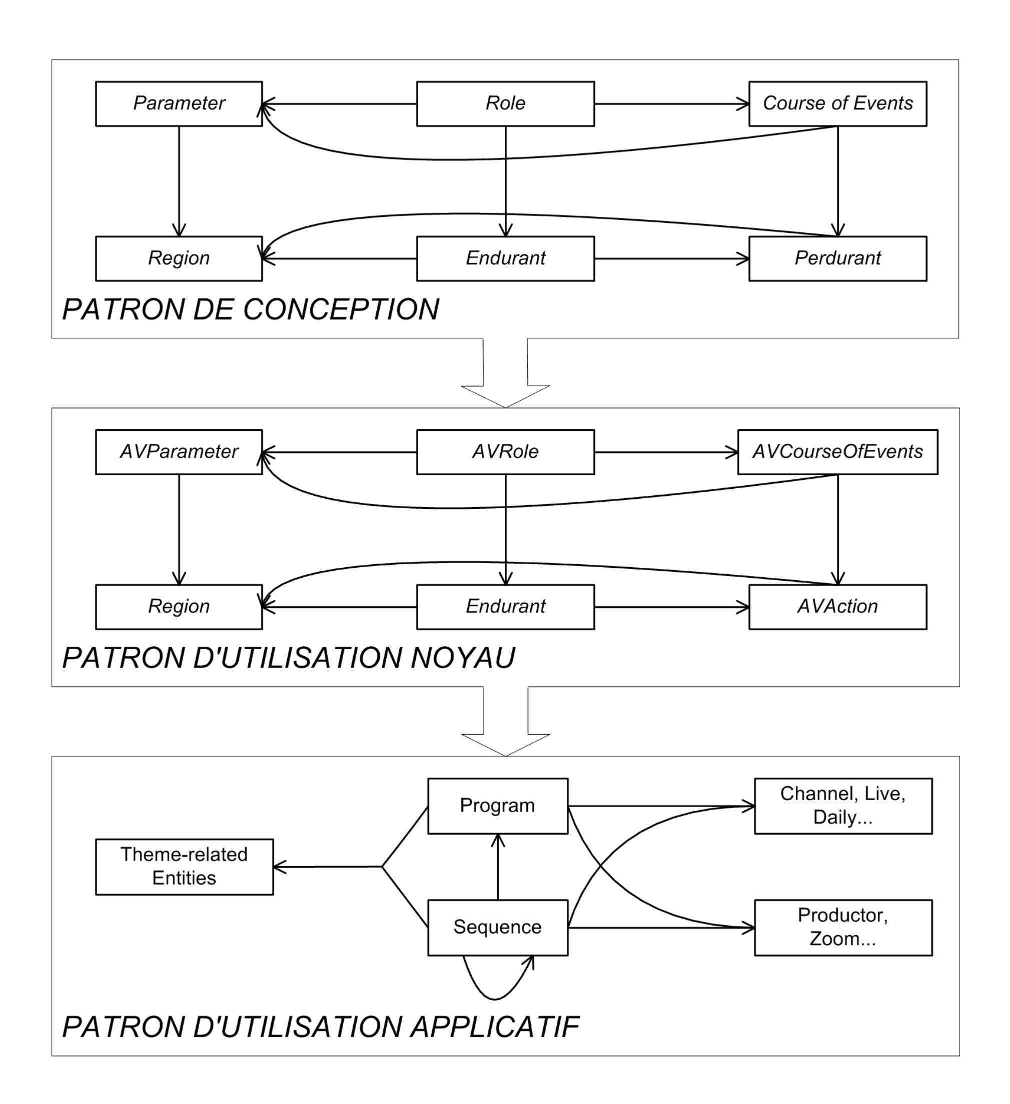


Fig. 4.8 Enchaînement des patrons ontologiques

Le patron noyau, lui, résulte d'un *passage au domaine* par rapport au PC. Cependant, il ne doit y avoir guère plus que cela, puisqu'on reste au niveau de l'ontologie *noyau* du domaine. Concepts et relations ont donc toujours un parfum fonctionnel, même s'ils sont dorénavant ancrés dans le domaine. Si on reprend l'exemple de la petite enfance du chapitre 3, on a vu que le patron d'indexation utilisait des notions comme `mère`, `SoinHygiène` et `pratiquéPar`. Et que ces notions s'opposaient à `Personne`, `participant` et `Action`, les plus génériques de l'ontologie du domaine concerné[42]. Ce sont justement de telles notions qui composent les patrons noyau : ceux-ci donnent de fait les motifs les plus génériques généralement utilisés dans un domaine.

Ces trois niveaux sont présents plus ou moins explicitement dans les différents travaux sur l'introduction des patrons de conception ontologiques et leur spécialisation. Mais notre récapitulatif, s'il explicite ces réflexions, les place aussi distinctement dans l'optique de légitimation applicative entreprise par le recours systématique à des patrons d'utilisation applicatifs. De fait, une proposition comme celle de [GCB04] montre bien nos trois niveaux – donnés indirectement dans les figures 4.4 et 4.5. Mais elle n'atteint pas le niveau de pertinence requis par les patrons applicatifs, puisque la structure du niveau applicatif est la même que celui du niveau noyau, elle-même reprise du patron de conception.

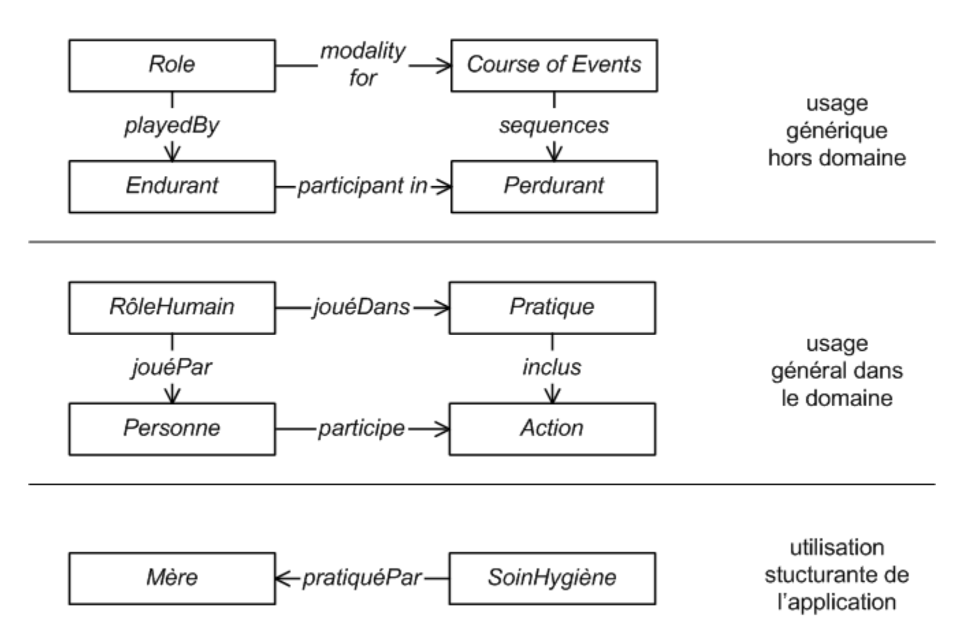


FIG. 4.9 Enchaînement des patrons ontologiques (partiels) pour la petite enfance

La figure 4.9 illustre bien ces différences. Nous montrons ici comment une partie du patron D&S[43], présentant comment un objet physique peut participer à un objet temporel, en jouant un rôle dans une séquence d'événements qui inclut cette activité. Dans le domaine, le patron noyau concernerait donc des personnes qui jouent des rôles particuliers, liées à la petite enfance, dans des actions précises, que l'on regroupe du point de vue de pratiques unificatrices. Une telle complexité est bien entendue exclue du patron d'indexation couramment utilisé, où les notions, plus spécialisées, ne sont pas organisées de la même façon : les pratiques existent, mais ne sont

[42]Dans le cas des graphes conceptuels, on se retrouve à un niveau de granularité descriptive comparable à celui qui est demandé pour la réalisation des graphes canoniques (cf. page 57) d'un domaine.

[43]Celui-ci peut encore une fois être utilisé dans un de nos domaines d'application. Il faut reconnaître que l'usage qu'il vise, la description, est évidemment proche de nos activités d'indexation.

pas au cœur des indexations les plus courantes, et les rôles les plus typiques, comme `Mère`, directement incorporés aux types d'individus qui les jouent.

Articuler patrons applicatifs et patrons noyaux par des connaissances de raisonnement

Si on veut pouvoir bénéficier des avantages de chacune des modélisations proposées par les patrons mobilisés lors de la conception de l'ontologie, que ce soient les patrons d'utilisation noyaux ou les patrons d'utilisation applicatifs, il faut introduire dans ce processus des stratégies de passage d'une structure à l'autre.

Nous proposons, pour faire coexister patrons génériques (patron de conception et patron noyau) et patron applicatif au sein d'un cadre méthodologique cohérent, d'articuler les deux formes d'expression par des connaissances de raisonnement formelles. L'objectif est d'autoriser un système d'inférence à passer de l'une à l'autre à chaque fois que cela est faisable. En particulier, il faudra présenter les liens « simples » demandés par le patron d'utilisation comme autant de *raccourcis relationnels* de chemins présents dans le patron de haut niveau.

Par exemple, pour notre ontologie de l'audiovisuel, on peut considérer qu'une relation `wasBroadcastedAt` entre un programme et une date est utile si l'on ne veut pas dire que le programme joue le rôle de message dans une séquence d'événements qui admet pour paramètre une date de diffusion valuée par ladite date. Il nous faut alors introduire un axiome (cf. figure 4.10) qui permettra de gérer simultanément les concepts et les relations des deux points de vue[44].

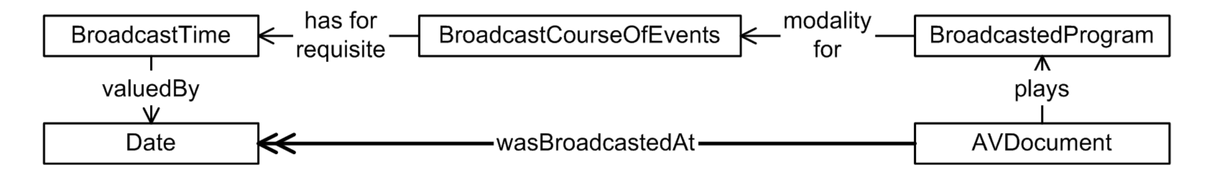


FIG. 4.10 Déduction d'une relation à partir d'assertions respectant le patron D&S

La double flèche indique la connaissance relationnelle déduite du chemin – complexe – initial.

Il est important de noter que si ces connaissances visent à aligner la structure du patron applicatif sur celle du patron de conception, le niveau auquel elles peuvent être données peut être plus général, dans la limite définie par le patron noyau. Par exemple, une relation `pratiquéPar`, dans l'ontologie de la petite enfance, pourrait être déduite de la connaissance selon laquelle une personne joue un rôle actif dans une pratique incluant l'action à laquelle on veut la relier (cf. figure 4.9). Il est évident que pour tenir compte des éventuelles variations légitimes autour du patron, mais aussi parce que le domaine est ainsi fait, on doit spécifier un tel axiome au niveau des concepts généraux de `Personne` et `Action`, et non `Mère` et `SoinHygiène`...

Ces connaissances de « traduction » seront ensuite opérationnalisées à l'aide de langages de représentation pour qu'elles puissent être exploitées par le SBC. Le code 4.2 donne la traduction de la règle de la figure 4.10 dans un des langages proposés pour représenter les règles dans le cadre du web sémantique, SWRL [HPB+04]. Il faut d'ailleurs remarquer qu'à ce stade on redevient tributaire de l'expressivité du langage retenu. Par exemple, dans SWRL, il ne sera pas possible

[44]On doit remarquer qu'une équivalence parfaite n'est pas forcément atteignable. Dans l'exemple donné, une information exprimée selon le patron de haut niveau contiendra des assertions qu'on ne pourra pas déduire précisément – il faudra recourir à la quantification existentielle – de connaissances issues du patron d'utilisation.

de représenter la contraposée de notre règle, puisque celle-ci fait appel à des individus introduits par des quantificateurs existentiels, individus auxquels on ne peut – dans les LD sur lesquelles s'appuient SWRL – faire référence pour construire des cycles tels que celui de la figure 4.10.

Il faut mentionner que la traduction d'une structure à une autre peut aussi passer par l'expression de définitions de concepts du patron noyau. Ainsi, de la définition par condition nécessaire et suffisante d'un *programme* (cf. figure 4.11), on peut déduire l'existence d'une séquence d'événements de diffusion dans laquelle ce programme joue le rôle d'objet diffusé, et réciproquement.

$$\texttt{Program} = \texttt{AVDocument} \sqcap (\exists\ \texttt{plays}\ (\texttt{BroadcastedProgram} \sqcap (\exists\ \texttt{modalityfor}\ \texttt{BroadcastCourseOfEvents})))$$

FIG. 4.11 Une définition possible pour le concept `Program`

Un programme est un document audiovisuel qui joue un rôle de programme diffusé dans une séquence d'événements de diffusion.

L'expression formelle et opérationnelle – donc exploitable automatiquement – de ces connaissances de raisonnement autorise un système à gérer simultanément un point de vue adéquat cognitivement et opérationnellement avec les usages du domaine, et un autre se référant plus à des principes de modélisation de haut niveau. Cette situation a deux avantages. D'abord, la connaissance est utilisable de manière satisfaisante pour l'application visée : on peut spécifier de façon naturelle des connaissances de raisonnement pertinentes, et on prescrit bien une description adaptée aux préoccupations applicatives. Ensuite, cette connaissance acquiert un statut plus consensuel : à chaque fois que c'était possible, on a tenté de se rapprocher de la structure du patron. Les informations de l'application peuvent être reprises dans d'autres applications construites autour du même noyau ontologique, ce qui ne peut qu'améliorer l'interopérabilité de ces systèmes.

4.4.4 Discussion

Après avoir constaté la pertinence de l'emploi des patrons d'indexation pour l'utilisation des ontologies, nous avons exploré le moyen d'en tenir compte dès le développement de celle-ci. De fait, nous ne sommes pas les seuls à avoir pris acte de l'existence de motifs réguliers dans les connaissances applicatives : se référant aux *design patterns* du génie logiciel, des travaux proposent déjà des solutions d'ingénierie ontologique qui exploitent des théories ou des entités abstraites, conçues pour être réutilisées dans des domaines plus précis. Après avoir parcouru ces solutions très diverses, nous avons vu que certaines étaient bien compatibles avec nos besoins. Des *patrons de conception ontologiques* mettent en relation des notions issues d'ontologies de haut niveau, de sorte que l'on puisse récupérer des notions, une structure et une axiomatisation pertinentes d'un point de vue théorique.

Les préoccupations qui ont dicté l'élaboration de tels patrons peuvent cependant être éloignées des besoins applicatifs tels qu'ils sont formalisés dans les patrons d'indexation. Une fois déterminé le patron d'utilisation applicatif, il faut le rattacher explicitement à la structure du patron de conception par l'intermédiaire de la relation d'héritage et de connaissances de raisonnement s'efforçant de réaliser des rapprochements entre structures différentes. A ce stade, le lecteur attentif pourra soulever la question de la place de la création du patron applicatif dans le processus de développement des ressources ontologiques : précède-t-il la conception de l'ontologie

```
<ruleml:imp>
  <ruleml:_body>
    <swrlx:classAtom>
      <owlx:Class owlx:name="Program" />
      <ruleml:var>prgm</ruleml:var>
    </swrlx:classAtom>
    <swrlx:classAtom>
      <owlx:Class owlx:name="BroadcastedProgram" />
      <ruleml:var>bcPrgm</ruleml:var>
    </swrlx:classAtom>
    <swrlx:classAtom>
      <owlx:Class owlx:name="BroadcastCourseOfEvents" />
      <ruleml:var>bcCOE</ruleml:var>
    </swrlx:classAtom>
    <swrlx:classAtom>
      <owlx:Class owlx:name="BroadcastTime" />
      <ruleml:var>bcTime</ruleml:var>
    </swrlx:classAtom>
    <swrlx:classAtom>
      <owlx:Class owlx:name="Date" />
      <ruleml:var>date</ruleml:var>
    </swrlx:classAtom>
    <swrlx:individualPropertyAtom  swrlx:property="&dolce;plays">
      <ruleml:var>prgm</ruleml:var>
      <ruleml:var>bcPrgm</ruleml:var>
    </swrlx:individualPropertyAtom>
    <swrlx:individualPropertyAtom  swrlx:property="&dolce;modality-for">
      <ruleml:var>bcPrgm</ruleml:var>
      <ruleml:var>bcCOE</ruleml:var>
    </swrlx:individualPropertyAtom>
    <swrlx:individualPropertyAtom  swrlx:property="&dolce;has-for-requisite">
      <ruleml:var>bcCOE</ruleml:var>
      <ruleml:var>bcTime</ruleml:var>
    </swrlx:individualPropertyAtom>
    <swrlx:individualPropertyAtom  swrlx:property="&dolce;valued-by">
      <ruleml:var>bcTime</ruleml:var>
      <ruleml:var>date</ruleml:var>
    </swrlx:individualPropertyAtom>
  </ruleml:_body>
  <ruleml:_head>
    <swrlx:individualPropertyAtom  swrlx:property="wasBroadcastedAt">
      <ruleml:var>prgm</ruleml:var>
      <ruleml:var>date</ruleml:var>
    </swrlx:individualPropertyAtom>
  </ruleml:_head>
</ruleml:imp>
```

Code 4.2 – Représentation en SWRL de la règle de la figure 4.10

applicative ? Ou au contraire son lien de spécialisation avec le patron de conception indique-t-il qu'on l'obtient après cette spécialisation ?

De fait, il n'y a pas de relation causale entre d'une part les connaissances fondamentales et d'autre part les connaissances ontologiques applicatives (notions, connaissances de raisonnement, patrons applicatifs) qui soit valable dans tous les cas de conception d'ontologies. On ne peut en effet dire que les connaissances que celle-ci contient ne peuvent avoir été obtenues qu'après considération des ressources de haut niveau. Dans l'approche que nous avons retenue, en l'occurrence, l'obtention des notions structurantes, issues des pratiques métier du domaine, peut précéder leur normalisation et leur formalisation, mais ce sont ces dernières qui achèvent d'en faire des ressources exploitables pour un SBC. De même, les patrons applicatifs pré-existeront souvent à leur rattachement aux patrons de conception. Mais le fait que soit disponibles des ressources déjà structurées et légitimées, pouvant servir de référentiel à la modélisation du domaine peut influencer leur création, même si le lien est moins fort que dans le cas du patron noyau. L'étape de conception des ressources de haut niveau est donc logiquement antérieure à celle de conception des ressources métier, mais on ne peut faire d'hypothèse sur la relation chronologique entre l'instant où ces dernières sont définies et celui où on les rattache aux premières[45].

Ce flou dans l'organisation du processus de spécialisation des patrons, ajouté aux contraintes temporelles de notre thèse, nous ont empêché d'ajouter une contribution technique aux solutions méthodologiques que nous venons de discuter. De surcroît, les opérations élémentaires d'adaptation du patron au niveau des notions (spécialisation, création de connaissances de raisonnement spécialisées) sont gérées par les outils classiques de développement d'ontologies. Ces mêmes outils proposent d'ailleurs des outils de visualisation d'ontologies[46] (ou d'extraits d'ontologies) qui seraient tout à fait adaptés à la manipulation graphique des patrons de conception, des patrons noyaux les spécialisant et des patrons applicatifs.

Néanmoins, il faut nuancer ce propos par l'observation du fait que les patrons, et notamment les patrons d'utilisation applicatifs, n'ont pas forcément de véritable légitimité *ontologique* au sens strict du terme. Il sont en effet définis à un niveau dont la pertinence, d'essence pragmatique, ne tient qu'à l'application envisagée, et non à une quelconque propriété des notions qui y apparaissent. Ainsi, les concepts du patron ne correspondent pas aux domaines des relations qui les lient : ces domaines, relevant plutôt du niveau *noyau*, sont souvent plus généraux. Or les éditeurs graphiques d'ontologies, non sans raisons, ne prennent en compte que les domaines des relations dans leur rendu graphique – ou non, d'ailleurs. Ces outils devraient donc inclure de véritables fonctionnalités de gestion de patrons, *via* la mise en avant explicite dans les éditeurs d'un niveau structurant, et la possibilité de travailler à partir de ce niveau et non plus seulement des niveaux génériques. En particulier, il faudrait veiller à :

- introduire dans le modèle ontologique géré par l'éditeur des « préférences » applicatives, comme la liste des concepts et relations présents dans le patron d'utilisation applicatif ;
- créer une interface de représentation de ce patron, permettant un visionnage de celui-ci en association avec les spécifications ontologiques – de facture « classique » – l'articulant avec les concepts et relations des deux autres patrons[47].

[45] Même dans l'approche de Gangemi et ses collègues, on peut se retrouver dans une situation où l'ontologie de domaine a été créée avant d'être rattachée au patron de conception de haut niveau, ce qui n'enlève rien au bénéfice de ce rattachement, qui prend alors le nom d'*alignement* [MGOS04].

[46] **Protégé** en propose plusieurs, disponibles à `http://protege.stanford.edu`.

[47] Il faudrait alors appliquer le même traitement de visualisation à ces deux patrons. Le patron de conception, en particulier, peut être considéré comme étant lui aussi au niveau « structurant », mais en comparaison des autres notions de l'ontologie fondamentale dont il est issu, ce qui le place par rapport à cette dernière dans la même position que le patron applicatif par rapport à l'ontologie de domaine.

Une telle instrumentation ne doit pas exiger un travail insurmontable. Malheureusement il n'a pas pu, faute de temps, être réalisé dans le cadre de notre thèse.

4.5 Conclusion

Ce chapitre s'est donc attaché à mieux prendre en compte les besoins liés à la création et l'utilisation des connaissances assertionnelles – pour nous, les index – lors du développement des ontologies. Poursuivant sur la lancée du chapitre 3, nous avons voulu voir comment on pouvait assister la création de substances ontologiques et de formes relationnelles adaptées à nos applications. Après avoir constaté que les méthodologies et outils existant ne traitaient pas complètement ces problèmes, nous avons proposé nos propres solutions, à la fois méthodologiques et techniques.

En ce qui concerne la création d'ontologies dont le vocabulaire serait spécifié de manière pertinente au regard de l'application visée, nous nous sommes tourné vers la méthodologie de Bruno Bachimont, qui introduit une phase de normalisation sémantique établissant des notions définies à l'aide des principes différentiels évoqués en section 3.2.2. Nous avons instrumenté cette méthode en réalisant un outil d'édition articulable avec les principaux environnements de la communauté, **DOE**.

S'agissant de l'obtention d'ontologies qui prennent convenablement en compte nos propositions en matière d'assistance à la structuration des index, à savoir les patrons d'indexation, nous avons observé qu'il était possible de reprendre des solutions à base de patrons de conception ontologiques. Nous avons repris la démarche d'Aldo Gangemi, qui consiste à utiliser des patrons de conception définis au niveau des ontologies fondamentales, en fonction d'usages génériques. Toutefois, nous avons remarqué que leur spécialisation directe est parfois insuffisante si l'on veut capturer dans toute leur finesse les besoins des applications. Pour résoudre ce problème, nous avons proposé d'introduire une articulation explicite entre patrons de conception et patrons d'utilisation applicatifs, articulation qui serait exploitable par le SBC utilisant l'ontologie ainsi conçue.

Au terme de cet effort méthodologique, nous parvenons donc à faire bénéficier le processus de conception d'ontologie d'une double légitimité. Tout d'abord, la conception prend bien en compte les pratiques – et les compréhensions – rencontrées dans le domaine d'application. On utilise la langue et ses verbalisations « métier » comme fondement même de la construction des concepts et des relations, et l'on se réfère à des patrons d'utilisation qui reflètent fidèlement les besoins de l'application en matière de description conceptuelle. Les solutions suggérées apportent donc à l'ensemble du processus un bien-fondé certain en ce qui concerne les *usages*. Ensuite, on bénéficie d'une légitimité et d'une cohérence *théoriques* importantes. L'expression naturelle des significations obéit à des principes justifiés par une théorie linguistique reconnue. Et les conséquences en termes de rapports hiérarchiques de ces prescriptions interprétatives sont respectées lors des étapes suivantes de formalisation et d'opérationnalisation. De plus, au cours de la conception, le rattachement des patrons applicatifs aux patrons de conception permet de situer l'engagement ontologique du concepteur par rapport à une théorie de haut niveau que l'on peut considérer comme une source de vraisemblance et de consensus.

En d'autres termes, on peut dire que nos propositions répondent à la fois aux problèmes d'utilisation et de reprise de l'ontologie. Car une plus grande réutilisabilité découle naturellement de la plus grande capacité de l'ontologie à faire consensus, et donc à être partagée, mais aussi de la possibilité qu'elle a de faciliter d'éventuels échanges entre systèmes partageant des éléments

de conceptualisation fondamentaux communs.

Qui plus est, les principes que nous recommandons sont de nature prescriptive : plutôt que de proposer des indications abstraites où l'utilisateur reste seul face à la spécification du contenu ontologique et la façon d'en organiser[48] les éléments, on indique des manières d'exploiter des informations concrètes et accessibles : expressions en langue, patrons de conception disponibles dans des bibliothèques d'ontologies fondamentales. Ces principes viennent donc compléter utilement les autres démarches d'ingénierie ontologique, attachées soit à l'organisation générale du cycle de développement, soit à la validation de son résultat selon des critères au champ d'application extrêmement précis.

Les solutions méthodologiques de conception que nous avons avancées dans ce chapitre l'ont été en complet accord avec les préconisations d'utilisation des ontologies dégagées en matière d'indexation dans le chapitre 3. De fait, l'ensemble de nos propositions forme à présent un cadre méthodologique cohérent et pertinent au regard des trois points que nous avions dégagés en introduction de cette partie. La conception du vocabulaire ontologique est en effet correctement guidée pour faciliter la compréhension de descriptions – point **M1**. Qui plus est, cette conception avec comme objectif explicite l'obtention des structures qui faciliteront la création des descriptions – point **M2** – ainsi que la création d'une forme de raisonnement qui soit pertinente pour le fonctionnement du SBC.

Il va nous falloir à présent refaire un bilan complet de ces propositions, dépassant celui qui a été fait par exemple dans notre article [IBL05] qui ne concernait que le volet des patrons d'utilisation et de conception, et surtout montrer qu'elles sont applicables. Nous allons pour cela faire une synthèse des expérimentations qui ont été menées au cours de cette thèse, et dont sont issus la majorité des exemples que l'on a vus dans ce manuscrit. Comment a-t-on pu appliquer nos solutions dans ces cas d'utilisation concrets ? Dans quelle mesure ces expérimentations valident-elles la justesse de ces propositions ? Quels sont les problèmes qui restent à traiter ?

[48] Que cette organisation repose sur l'emploi de la relation fondamentale de subsomption ou sur des agencements privilégiés entre concepts et relations conceptuelles.

Chapitre 5

Expérimentations et discussions

5.1 Récapitulatif des apports méthodologiques de cette thèse

La première des contributions de ce manuscrit est une réflexion sur les besoins concrets que posent les systèmes d'indexation et de recherche audiovisuelle. Nous avons en effet constaté qu'indexer est une tâche essentiellement interprétative d'analyse et de reformulation qui demande beaucoup de soin et de savoir-faire. Et que la recherche dans une base d'index doit tenir compte de toutes les « imprécisions » qui pourraient apparaître dans ceux-ci. En particulier, nombre d'informations sont nécessairement laissées implicites par l'indexeur, qui fait confiance aux connaissances que le recherchiste, dans le contexte bien précis qui encadre chaque pratique d'indexation, est supposé partager avec lui. Le problème est donc qu'une certaine *continuité sémantique* doit être conservée, des pratiques et compréhensions des indexeurs à celles des utilisateurs qui accèdent au résultat de leur activité.

L'état de l'art recommande pour contrôler la validité de cette continuité sémantique le recours à des *langages documentaires*. Ces langages permettent de contrôler les expressions qui servent d'index, mais aussi de fournir des référentiels qui guident l'interprétation du vocabulaire qu'ils fournissent. Les *thesauri* sont les plus élaborés de ces référentiels documentaires, qui présentent des termes normés dans un contexte structuré par des liens de spécialisation sémantique. Cependant, ces thesauri ne suffisent pas à garantir un niveau d'efficacité satisfaisant en ce qui concerne la prise en charge de la continuité sémantique par le système lui-même. Les points d'amélioration, on l'a vu, concernent :

– l'expressivité autorisée pour les index : les contenus documentaires sont complexes, spécialement dans le cas des documents audiovisuels. Il faut introduire un vocabulaire conceptuel riche et adapté à l'application envisagée. Il faut aussi pouvoir créer des index liant explicitement ces éléments conceptuels à l'aide de notions relationnelles qui soient spécifiques à l'application concernée par l'indexation ;
– le contrôle : le référentiel sémantique du langage doit être défini de manière plus rigoureuse que dans les thesauri. Cela impose deux efforts de spécification : l'un, en direction des utilisateurs, dont les interprétations, intrinséquement informelles et naturelles, doivent être guidées de façon suffisamment précise, et l'autre, formel et donc artificiel, pour le système, qui vérifiera la validité sémantique des index produits ;
– la manipulation des descriptions : le système peut prendre en charge les tâches les plus simples de ré-interprétation et de complétion que les utilisateurs accomplissent lorsqu'ils sont confrontés aux index lors de leurs recherches.

Les solutions de représentation de connaissances, et en particulier celles qui utilisent les *ontologies*, peuvent être adaptées au cas de l'indexation. Les ontologies fournissent les primitives de connaissances qui vont constituer un langage de représentation permettant de créer des index pertinents dans le cadre applicatif visé. Ces primitives comportent des *concepts* mais aussi des *relations* qui permettent de construire des index structurés. De surcroît, les notions ontologiques voient leurs significations détaillées de sorte à ce qu'elles puissent être interprétées suivant des critères formels stricts, prélude à leur manipulation directe par un système à base de connaissances. Finalement, les ontologies sont encodables dans des méta-langages de représentation de connaissances, qui permettent de définir des langages de représentation d'index dont les primitives, liées à des connaissances de raisonnement, sont accessibles par les SBC.

Nous avons vu tout au long du second chapitre, au travers d'exemples issus de nos expérimentations, et en particulier de celles réalisées au cours du projet OPALES, comment les fonctionnalités habituellement offertes par les systèmes ontologiques répondaient à nos besoins en termes de conception de référentiels sémantiques formels : création de hiérarchies à la signification précise, définitions formelles des concepts, règles de composition relationnelles... On peut dès lors envisager des procédures de contrôle des expressions construites, ainsi que l'utilisation de mécanismes d'inférence exploitant ces significations pour compléter les informations fournies par les index et mieux répondre aux requêtes posées par les utilisateurs. En d'autres termes, un système qui fasse bénéficier d'une meilleure continuité sémantique d'un bout à l'autre de la chaîne documentaire.

Le problème est que de telles solutions ne sont pas aisées à mettre en place. Le cadre dans lequel nous nous sommes situés est celui d'une indexation manuelle, où l'interprétation de l'indexeur joue un rôle crucial. Et l'utilisation pour une telle indexation de primitives à la signification formelle complexifie les index, alors que ceux-ci, même dans un contexte plus « naturel », sont délicats à concevoir et comprendre. Nous avons donc mis en place des procédés facilitant les rapports de l'utilisateur avec les deux aspects fondamentaux des index, à savoir leur *substance* – le vocabulaire qu'ils emploient – et leur *forme* – leur structure relationnelle.

Pour la première, nous avons repris la méthode proposée par Bruno Bachimont, qui rattache les primitives du langage d'indexation à des signifiés « métier » formulés de façon rationnelle à l'aide de la langue : en exprimant les similarités et les différences entre notions à l'aide de principes différentiels, on leur donne une signification naturelle mais précise. Pour la seconde, nous avons détaillé la manière d'utiliser des *patrons d'indexation* qui reflètent les besoins de l'application en matière de contenu des index, comme des formulaires peuvent le faire dans les systèmes classiques. Ces graphes sont des structures comprenant les notions qui *structurent* le champ de l'application, et on peut s'en servir comme point de départ des indexations effectives, en les adaptant aux cas concrets rencontrés. En concevant des connaissances de raisonnement qui soient dédiées à la déduction de connaissances permettant de passer des éléments du patron à d'autres structures ou réciproquement, on peut gérer un certain degré de variations par rapport à ce patron, ce qui permet de considérer celui-ci comme un *canon* – en admettant qu'un canon puisse être dérivé en des index qui ne sont pas *stricto sensu* des spécialisations – pour le SBC, canon qui est certes implicite, mais rendu pleinement opérationnel par le recours aux inférences qui lui sont liées.

Ces deux méthodes aident donc à construire des descriptions légitimes par rapport aux pratiques ciblées : leur vocabulaire se place dans un contexte interprétatif qui se rapporte clairement à ce qui se passe dans le contexte applicatif, et leur contenu se conçoit explicitement en référence à des structures qui capitalisent l'expertise en matière d'indexation dans ce même contexte

applicatif. *In fine*, on construit un système d'assistance à l'indexation qui, couplé à des mécanismes d'inférence appliquant des règles de déduction en accord avec la signification de tous ces éléments, favorise l'obtention d'un niveau élevé de continuité sémantique.

Mais aux difficultés de la création, de l'utilisation et de l'accès aux index ontologiques vient s'ajouter celles de la conception des ressources à partir desquelles sont construits ces index. Les ontologies sont des artefacts dont le développement est délicat : les connaissances à encoder sont complexes, surtout dans l'optique que nous avons retenue, qui rend nécessaire des spécifications riches. Qui plus est, il faut garantir leur pertinence vis-à-vis des applications pour lesquelles sont conçus les SBC qui emploient ces ontologies.

Nous avons vu au cours du chapitre 4 que loin de ne constituer qu'une source de complexité supplémentaire, la considération des problèmes métiers dès la conception permet de rationaliser ce processus et d'assister le concepteur dans la détermination d'un contenu ontologique pertinent pour l'application. Et que ceci se fait d'une manière avantageusement complémentaire des autres propositions méthodologiques du domaine de l'ingénierie ontologique, qui se concentrent sur le cycle de vie général de l'ontologie, sur l'obtention de proto-ontologies pour lesquelles un important travail de conceptualisation reste à faire, ou bien encore sur la validation formelle d'ontologies ayant déjà bénéficié d'un engagement significatif de la part de leur concepteur.

Tout d'abord, la méthodologie de Bruno Bachimont va bien au-delà de l'assignation d'une signification métier naturelle aux primitives. Cette assignation est en effet le fruit d'une *normalisation* sémantique qui débouche sur la création de hiérarchies de concepts et de relations directement utilisables pour la création d'une ontologie de primitives dont la signification est formalisée mais non encodée dans un langage opérationnel particulier. On opérationnalise ensuite cette formalisation – de la manière la plus automatisée possible – pour rendre l'ontologie exploitable par un SBC. Tout en réfléchissant sur des problèmes méthodologiques concrets soulevés par cette approche, nous avons conçu un outil d'édition d'ontologies qui en plus de la normalisation sémantique autorise une formalisation et une opérationnalisation limitées aux capacités d'un certain type de supports de graphes conceptuels. En accord avec la méthodologie adaptée, cet outil offre tout de même la possibilité, *via* des mécanismes d'export utilisant les langages de représentation standards, de poursuivre ces deux dernières étapes dans des environnements qui, comme on l'a montré, se concentrent exclusivement sur ces aspects et sont donc plus appropriés.

Ensuite, les patrons d'indexation peuvent s'insérer tout naturellement dans le processus de construction de l'ontologie, en tant que *patrons d'utilisation* de ses concepts et relations. Ils permettent de donner une justification pratique à des approches de réutilisation de *patrons de conception*. En effet, il est possible de rattacher le développement d'une ontologie de domaine à des patrons de conception qui prélèvent au sein d'ontologies de haut niveau les notions utiles à un usage qui est ciblé mais nécessairement indépendant d'un domaine, et les présentent sous forme d'un motif relationnel adapté à cet usage. Ces patrons de conception permettent de réutiliser des contenus ontologiques dans le processus de conception des ontologies de domaine et d'application, ce qui le facilite grandement. Ils permettent également d'obtenir des structures qui sont légitimées, mais uniquement du point de vue de la consistance et de la vraisemblance théorique. L'agencement obtenu en les spécialisant simplement, en particulier, peut ne pas suffire à rendre compte des usages concrets rencontrés dans un domaine. En construisant une articulation explicite – employant spécialisation et connaissance de raisonnement – entre les patrons de conception et les patrons d'utilisation, plus pragmatiques, on permet aux SBC de bénéficier en même temps d'une validation théorique et d'une validation applicative.

Un tel cadre méthodologique permet, dans le contexte de l'indexation à base de connaissances, une conception et une utilisation des ontologies qui se fassent plus facilement, et de façon mieux adaptée aux besoins que nous avons dégagés. De fait, comme les nombreux exemples concrets dont nous avons illustrés nos propositions le laissent envisager, ce cadre n'a pas été élaboré *in vitro*, uniquement à partir de l'état de l'art – certes important – et d'un effort théorique de rationalisation méthodologique. Il est en effet également le fruit d'expérimentations à différentes échelles et sur différents sujets – tous cependant liés à l'indexation de documents audiovisuels. Ces expérimentations ont permis d'observer des problèmes concrets, mais aussi de tester, de faire mûrir les hypothèses ou les outils que nous avions avancés. Nous allons à présent détailler les plus significatives d'entre elles. Nous ajouterons à cette présentation l'exposé d'un certain nombre de points de discussion, qui pour les systèmes documentaires audiovisuels à base d'ontologies constituent autant de problèmes toujours à régler, ou bien des pistes qui nous semblent intéressantes à évoquer pour de futures recherches...

5.2 Expérimentations ontologiques

Pour chacune des expériences de conception d'ontologie dans lesquelles nous avons été impliqué, nous allons :

- présenter d'un point de vue qualitatif et quantitatif les concepts, les relations, les règles et graphes patrons qui constituent le *contenu* de l'ontologie ;
- évoquer la manière dont l'ontologie est exploitée par un SBC : son *usage* ;
- discuter de l'éclairage que leur conception et leur utilisation peut apporter sur nos hypothèses méthodologiques.

5.2.1 Cyclisme

Chronologiquement, la première des expérimentations de conception d'ontologie auxquelles nous ayons participé est celle de l'ontologie du « cyclisme ». Initiée par Estelle Le Roux et Raphaël Troncy pour les besoins de leur thèses respectives, cette ontologie a été conçue pour :

- demontrer la faisabilité d'une approche d'indexation articulant langages de structuration documentaire[1], description au niveau de la connaissance de la structure documentaire et description thématique [Tro04].
- servir de référentiel conceptuel à un outil d'extraction de connaissances – **SEIGO** – qui analyse un ensemble de documents textuels relatif à un corpus audiovisuel, pour produire des descriptions conceptuelles structurées candidates au statut d'index [Le 03].

Cette ontologie a pour objectif d'autoriser la description de faits identifiés – avec l'aide de documentalistes spécialistes du sujet à l'INA – comme pertinents pour le Tour de France cycliste, comme les arrivées d'étape, les classements, les abandons, etc.

Création de l'ontologie

Après l'extraction de libellés linguistiques des textes concernés et leur organisation en une première hiérarchie – phases auxquelles nous n'avons pas participé – nous avons appliqué la première étape de la méthodologie de Bruno Bachimont, la normalisation sémantique. On peut

[1] Ces langages sont prioritairement intéressés par l'encodage d'une description structurelle et fonctionnelle d'un document, la description du contenu thématique, employant un langage documentaire contrôlé, suivant l'acception qui a été la nôtre jusqu'à présent, ne venant qu'ensuite.

se rendre compte, à travers les exemples de la section 3.2.2 (page 89) et 4.3.3 (page 131) de la manière dont cette opération a été menée dans cette ontologie.

Il faut remarquer que pour rendre les niveaux les plus élevés de nos hiérarchies plus cohérents que ce que les définitions métier permettent d'obtenir, nous avons réutilisé la hiérarchie de concepts de haut niveau introduite dans l'ontologie utilisée dans le projet MENELAS. Cette hiérarchie a été construite en suivant la méthodologie ARCHONTE [BBCZ95], ce qui nous a permis de bénéficier immédiatement du travail de conceptualisation abstrait accompli alors. Un tel effort de rattachement permet, comme cela a été déjà évoqué auparavant à propos de la réutilisation des concepts de haut niveau de patrons de conception, de clarifier l'engagement ontologique tout en augmentant le potentiel de ré-utilisation de notions initialement conçues pour une application précise.

En ce qui concerne les relations, nous sommes partis de la typologie proposée dans les travaux du laboratoire LaLICC [Des87]. Mais nous avons dû faire de nombreux aménagements pragmatiques à cette typologie qui résultait d'une réflexion linguistique rigoureuse : le rattachement privilégié de certains *attributs* à des concepts particuliers peut pousser à regrouper davantage les relations suivant les objets auxquelles elles s'appliquent, avant même de considérer leur signification propre. La figure 5.1 montre en partie les notions résultant de cette adaptation de ressources de haut niveau.

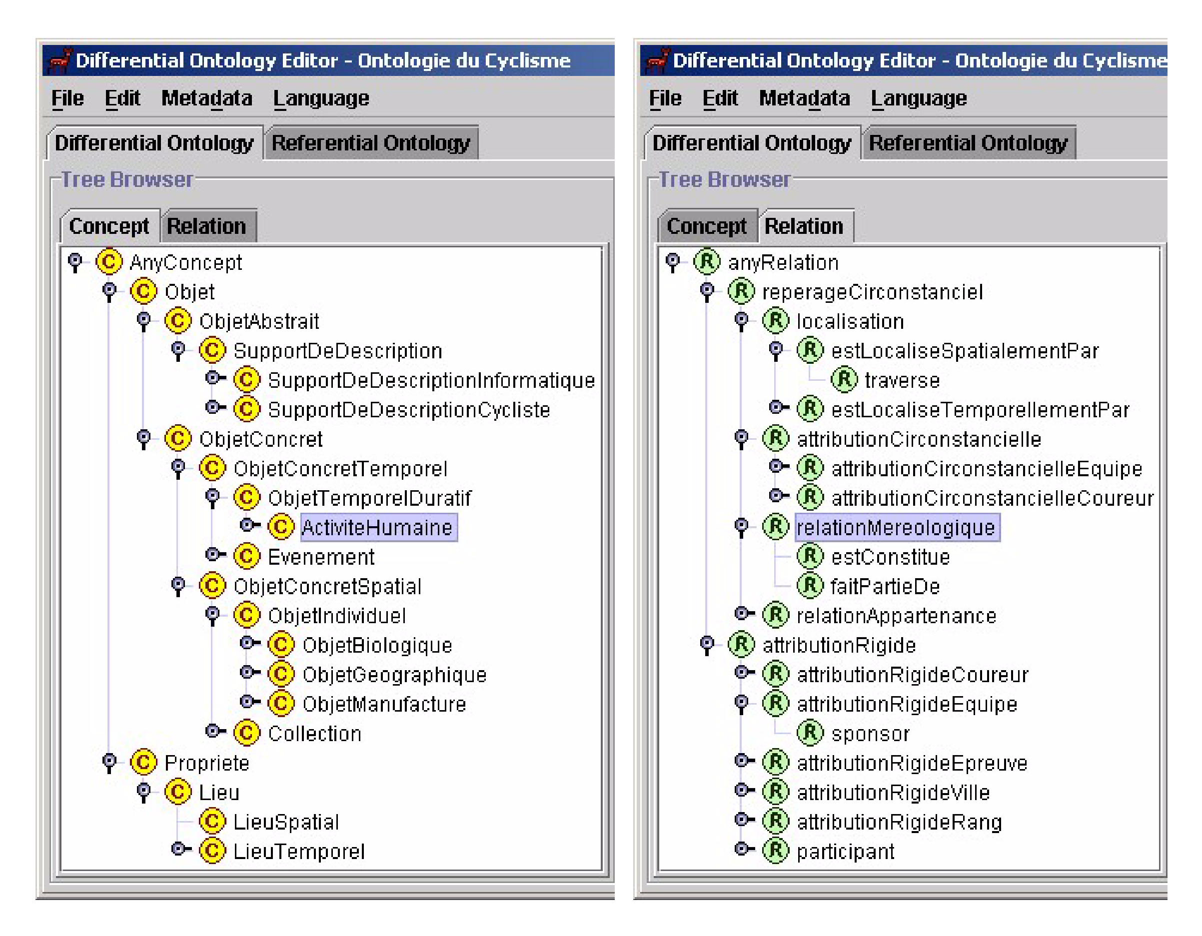


FIG. 5.1 Concepts et Relations de haut niveau de l'ontologie du cyclisme, extrait de [TI02]

L'ontologie a fait ensuite l'objet d'une formalisation et d'une opérationnalisation. Nous ne

nous attarderons pas sur les répercussions de ces deux dernières sur le contenu ontologique, puisque les pratiques ciblées ne nécessitaient pas l'introduction de connaissances de raisonnement complexes. En particulier, **SEIGO** ne demandait qu'un répertoire arborescent de concepts, avec les relations qui pouvaient tenir entre ceux-ci. Cette dernière condition a tout de même fait que certaines définitions ont été considérées. Par exemple, il pouvait être utile, pour l'extraction d'information, de savoir qu'une `Equipe` a *nécessairement* un `leader`. Dès que l'analyseur lexico-syntaxique reconnaît une occurrence d'une équipe, il peut déclencher les patrons d'analyse dont l'objectif est de reconnaître un coureur lié à cette équipe – le co-domaine de la relation `aCommeLeader` est `CoureurCycliste` – et créer la relation correspondante entre les deux. Certaines définitions de concepts parmi les plus simples ont donc été rajoutées à une formalisation élémentaire, ainsi qu'on peut le constater en consultant les exemples de la section 4.3.1 ou l'annexe C de la thèse de Raphaël Troncy, où une version OWL de cette ontologie est livrée dans son intégralité [Tro04]. En tout et pour tout, l'ontologie[2] du cyclisme compte 97 concepts et 61 relations, pour une profondeur maximale de 10.

Il est intéressant de mentionner ici que cette ontologie n'était pas destinée directement à l'indexation humaine, et ne comporte donc pas de patron d'indexation présenté explicitement comme répondant à la préoccupation globale d'une application. Pourtant, le besoin de déterminer des structures descriptives récurrentes a été présent, même si ces structures – multiples – se présentent sous l'aspect élémentaire de définitions – par condition nécessaire – introduisant les entités toujours liées à un concept *structurant* pour l'application. De fait, comme dans l'exemple de MENELAS que nous avons discuté en page 102, on a affaire à ici des patrons implicites et locaux, destinés à un usage d'extraction semi-automatique d'information.

Analyse ontologique formelle et cyclisme

Pour cette expérimentation, il faut également ajouter que dans la perspective de nos investigations méthodologiques, nous avons cherché à appliquer aux concepts de notre ontologie les principes de vérification *formelle* apportés par Nicola Guarino (cf. section 4.2.2).

La première de nos remarques, d'ordre pratique, est que cette expérience s'est faite de façon *ad hoc*, puisqu'au moment où nous l'avons conduite, il n'existait pas encore d'interface de saisie et de vérification de telles informations formelles, comme ODECLEAN [FG02]. Ensuite, d'un point du vue plus fondamental, et comme nous l'avons déjà indiqué dans la section 4.2.2, l'application des principes formels de Guarino pour les concepts n'a pas modifié de manière notable l'organisation que nous avions obtenue lors de la normalisation sémantique.

De fait, nous avons pu observer que la clarification obtenue suite à cette normalisation a pour effet de lever un certain nombre d'ambiguïtés, dont beaucoup comptent parmi celles que dénoncent Nicola Guarino et ses collègues. Par exemple, dans l'un des articles exposant leur méthodologie d'analyse ontologique [WG01], ceux-ci présentent le cas du concept `Pays`. Ce concept peut, dans une première approche, être modélisé comme une `LocalisationSpatiale`, ce qui traduit son ancrage géographique, mais aussi comme un `AgentLégal`, pour rendre compte de son rôle vis-à-vis des sociétés humaines. On serait donc dans une situation où il serait pertinent de recourir à l'héritage multiple pour définir cette notion. Pourtant, l'analyse ontologique de Welty et Guarino démontre qu'il s'agit là d'un cas typique de polysémie due au télescopage de deux regards relativement différents sur le monde. De fait, si le concept `Pays` était introduit de cette sorte, il devrait être *anti-rigide* d'un point de vue formel : n'importe quel pays peut cesser d'exister, sans que le lieu spatial sous-jacent disparaisse lui aussi. Le prétendu lien d'instanciation

[2]Celle-ci est disponible sur la partie du site d'OPALES hébergeant **DOE**, `http://opales.ina.fr/public/`.

entre le concept de pays et ce lieu est donc rompu : on peut par conséquent déduire que dans ce cas le concept `Pays` n'est pas essentiel pour toutes ses instances, et donc qu'il est anti-rigide. Ce qui est en contradiction avec l'idée qu'on se fait d'un pays : s'il peut changer de régime politique, ou de localisation, le fait est qu'il demeure dès que son existence a été affirmée une fois. Pour remédier à ce problème, il faut donc découpler les deux notions, et introduire deux concepts distincts de `RégionGéographique` comme `ObjetPhysique` et de `Pays` comme `ObjetSocial` et comme `LieuNonPhysique`, comme cela est fait dans l'ontologie Dolce [GGM+02].

Cependant, nous n'avons pas fait un tel choix dans notre ontologie. En effet, lorsque l'on indique que le Tour de France *traverse* un pays, il va de soi que l'on considère un objet spatial. On a là affaire à une figure de style, une métonymie où l'on utilise une entité pour désigner une autre à laquelle elle est liée. En l'occurrence, l'objet administratif est bien plus *structurant* pour le sens commun pour désigner le lieu géographique. Cela est d'autant plus valable pour le monde des courses cyclistes, qui joue toujours sur la contextualisation nationale[3] des coureurs ou des équipes. Le fait est que cette figure de style est devenue fondamentale pour la conceptualisation de l'application, qui ne s'attache plus le moins du monde aux subtilités administratives, sociales ou historiques qui motivent le découpage géographique qui a créé le lieu traversé : pour notre application, le pays est essentiellement rattaché à sa couverture géographique. Il s'agit d'une sorte de changement de type[4], mais ce glissement, parce que l'on se restreint à un monde bien particulier, est à présent considéré comme primordial pour la conceptualisation proposée.

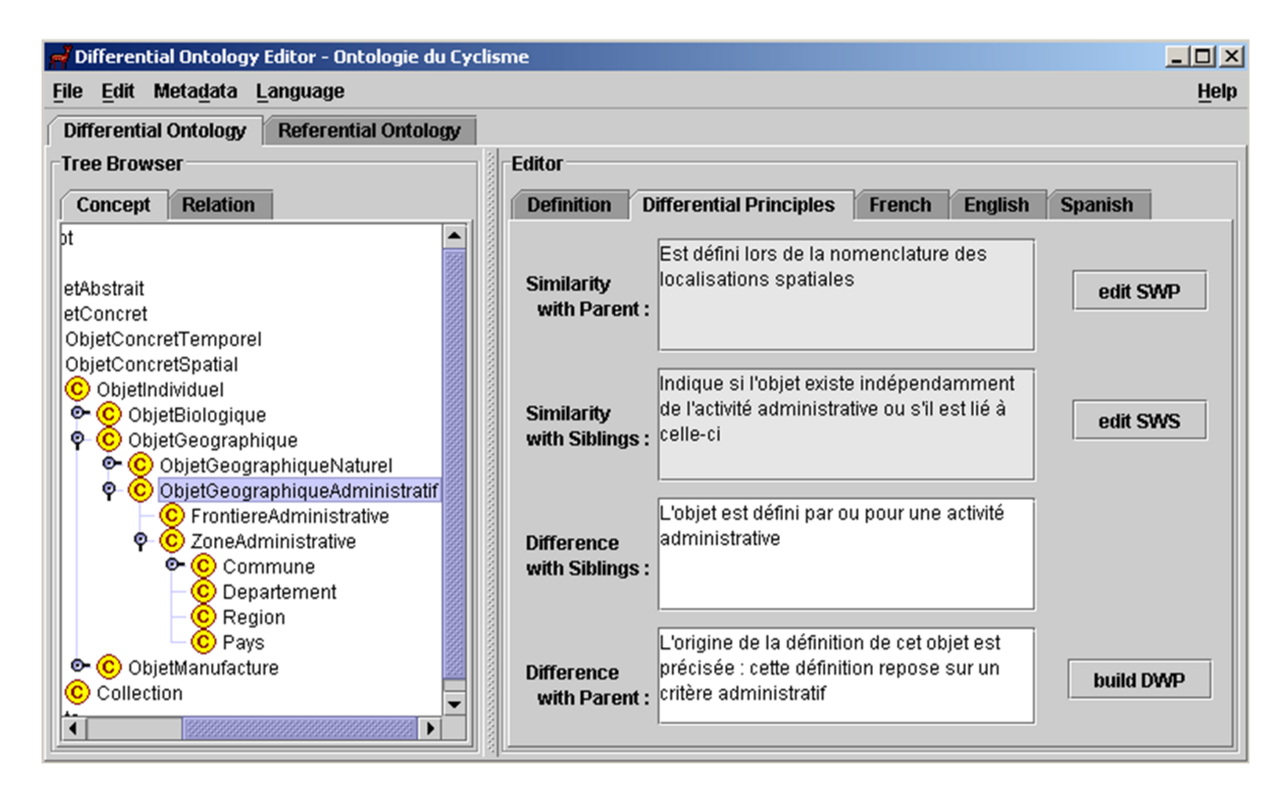


Fig. 5.2 Définition différentielle du concept `ObjetGeographiqueAdministratif`

[3]Ou plus locale, voir l'importance prise très souvent par le « régional de l'étape » . . .

[4]Des travaux du laboratoire lalicc, dans la lignée des constatations faites dans [Jou93], mentionnent l'existence d'opérateurs de changement du type sémantique d'une unité linguistique. Un *objet individuel* comme Paris – introduit comme capitale administrative – peut ainsi être considérée comme un *lieu spatial*, s'il est impliqué par exemple dans un *schème* faisant intervenir un mouvement – tel que pour « Jean va à Paris ».

Dès notre ontologie différentielle, nous avions donc introduit la notion d'`ObjetGéographique-Administratif`, comme un objet étant défini de manière privilégiée par des critères humains, et administratifs en particulier, mais existant essentiellement dans l'espace (voir figure 5.2). Et l'attribution des méta-propriétés formelles ne vient pas remettre en cause ce choix, puisqu'on se place dans un monde applicatif plus restreint que celui retenu par Welty et Guarino dans [WG01] : *le temps d'un Tour de France*, la couverture spatiale d'un `Pays` ne change pas, on peut donc considérer que ce concept est rigide même s'il est instancié par un lieu. Le fait que ce lien pourrait être remis en question dans des applications plus génériques n'est pas pertinent pour notre cadre applicatif limité[5]. Comme on peut le déduire[6] en observant le tableau 5.1, il n'y a alors aucune contre-indication à ce que les trois concepts d'`ObjetGéographique`, d'`ObjetGéographiqueAdministratif` et de `Pays` soient sur une même ligne de spécialisation.

`ObjetGéographique`	
rigidité :	+R
identité :	+I (condition nécessaire héritée d'`ObjetConcret` : deux objets géographiques sont identiques s'ils ont la même localisation spatio-temporelle)
unité :	+U (unité topologique : toutes les éventuelles parties de l'objet géographique sont localisées dans la zone spatio-temporelle qu'il couvre)
dépendance :	-D (on ne peut pas dire *a priori* que tous les objets géographiques dépendent d'objets instanciant un concept particulier)
`ObjetGéographiqueAdministratif`	
rigidité :	+R
identité :	+O (deux objets correspondant à la zone d'un même exercice administratif sont identiques)
unité :	+U (héritée d'`ObjetGéographique`)
dépendance :	+D (un objet n'existe que si une activité ou un niveau administratif précis ont été définis)
`Pays`	
rigidité :	+R
identité :	+O (deux pays ayant – au moins – un même gouvernement sont identiques)
unité :	+U (héritée d'`ObjetGéographique`)
dépendance :	+D (héritée d'`ObjetGéographiqueAdministratif`)

TAB. 5.1 Méta-propriétés associées à des concepts géographiques de l'ontologie du cyclisme

Pour chaque méta-propriété `X`, le symbole `+X` (respectivement `-X`) indique qu'elle est associée (respectivement qu'elle ne peut être associée) à la propriété considérée. `R`, `I`, `U` et `D` renvoient respectivement à la rigidité*, à l'*identité*, à l'*unité *et à la* dépendance*. La balise `+O` pour la méta-propriété d'identité indique que le concept fournit son propre critère d'identité, c'est-à-dire que celui-ci ne provient pas d'un concept plus général. Pour plus de détails, se reporter à [WG01] et à la section 4.2.2 en page 120.*

Comme on vient de l'évoquer, cet écart par rapport aux préconisations issues de l'ontologie formelle vient de ce que l'ancrage dans l'application supprime certains des effets de polysémie tant redoutés par Guarino, en limitant les *mondes possibles* à étudier pour attribuer les méta-

[5]Nous sommes néanmoins conscient que dans une optique plus large, où l'on voudrait par exemple *aligner* notre ontologie avec des ontologies dédiées à d'autres applications, cela pourrait engendrer des problèmes majeurs. Mais cette thèse n'avait pas à traiter de ces questions...

[6]Nous passons les détails liés aux autres méta-propriétés : en l'occurrence, nos choix en ce qui concernent celles-ci ne contredisent pas ce que [WG01] a proposé.

propriétés aux concepts de l'ontologie. On trouve un autre exemple relativement clair dans la branche de l'ontologie dédiée à la classification des personnes présentes sur le Tour. En effet, comme on peut le voir dans la figure 4.2, nous avons spécialisé le concept `Personne` en `Coureur-Cycliste`, `DirecteurEpreuve`, `MedecinEquipe`, etc. Or, pour Guarino, `Personne` est un concept rigide qui répond à tous les critères formels de définition d'un *type* [GW00a]. Or toutes les spécialisations que nous lui avons attachées sont anti-rigides : *dans l'absolu*, un coureur pourrait très bien être autre chose qu'un coureur – ce concept serait donc un *rôle*. Le fait est que *le temps d'un Tour de France*, les individus qui instancient `Coureur` sont *essentiellement* des coureurs cyclistes, quelle que soit la tournure que peuvent prendre les événements. Il est donc toujours pertinent de rattacher par un lien de subsomption les coureurs à la notion de personne.

L'analyse ontologique suivant les principes différentiels n'est donc pas vraiment contradictoire avec la méthode de « nettoyage » de Guarino, dans la mesure où l'expression linguistique est censée rendre compte d'une dimension cognitive dont relèvent également les mondes possibles utilisés par l'analyse ontologique formelle [Kri82]. Nous retrouvons, concrètement, l'analyse de Bruno Bachimont qui dans [Bac01] expliquait que les deux méthodes, si elles avaient des modalités d'action différentes – l'une effectuant un nettoyage *a posteriori*, l'autre construisant plutôt que filtrant un existant –, répondaient toutes les deux à une objectif de contrôle. De fait, dans le cas de notre coureur cycliste, nous n'avons pas renoncé à l'analyse formelle, nous l'avons seulement *située* dans une application bien particulière en limitant les mondes possibles à prendre en compte. *In fine*, cette analyse permet toujours de confirmer par des moyens formels – et *formalisés*[7] – un engagement ontologique « métier ». Cela est d'autant plus intéressant, au niveau de cet engagement ontologique contracté, si les choix de modélisation, qui dépendent nécessairement dans notre cas d'une application précise, se font à l'encontre de ceux qui ont été faits pour des ontologies de plus haut niveau.

5.2.2 Opales

Opales a été le premier et le plus important des véritables cadres applicatifs pour notre thèse. Pour rappel, le projet visait à la description de documents audiovisuels relatifs à deux points de vue particuliers : la « petite enfance » et l'« eau ».

Le premier concernait des documents à vocation ethnologique – 13 documents pour à peu près 6 heures de vidéo. La stratégie de description, en accord avec les experts de ce domaine participant au projet (documentalistes, chercheurs), s'est concentrée sur les actions de soins prodiguées aux nouveaux-nés : toilette, allaitement, endormissement. Avec Véronique Malaisé, nous avons défini une ontologie, permettant de traiter ces aspects. Véronique Malaisé a tout d'abord travaillé sur l'extraction de termes à partir de ressources textuelles ou terminologiques du domaine. Ensemble, nous avons ensuite appliqué la normalisation sémantique préconisée par Bruno Bachimont, à la lumière du contexte textuel dans lequel les termes apparaissaient. Il faut également remarquer que, comme dans le cas de l'ontologie du cyclisme, nous nous sommes tournés vers les concepts de haut niveau de l'ontologie du projet Menelas pour organiser le haut niveau de notre propre ontologie.

C'est à ce moment que nous avons pu déterminer le patron d'indexation tel qu'illustré dans la figure 3.10 de la page 99 : des mères décrites par leur posture pratiquent des soins d'hygiène sur

[7] L'assignation des méta-propriétés relevant du niveau de la forme des concepts assigne en effet à ceux-ci des lois dans le cadre d'une logique modale. Par exemple, pour le concept `Personne`, et si l'on renonce à considérer l'indexation temporelle de tous les prédicats qu'introduisent Guarino et ses collègues, on a $\Box \forall x\,(\texttt{Personne}(x) \rightarrow \Box\,\texttt{Personne}(x))$, où $\Box$ est le quantificateur modal de la *nécessité*.

leur bébé, actions qui peuvent être repérées spatialement par différents éléments de localisation.

Au cours de la formalisation de l'ontologie, nous avons pu rajouter des notions qui en améliorent la lisibilité, notamment des concepts liés à une fonction particulière dans le graphe patron. Ces concepts regroupent des entités que leur définition différentielle rendait incompatible : par exemple, le concept `Protagoniste` permet de donner une généralisation commune aux concepts de `Personne` et de `GroupeHumain` qui héritent de notions différentielles en opposition, et ce à un niveau qui fait toujours sens du point de vue métier. Ces concepts sont intégrés de manière naturelle dans la structure ontologique référentielle issue de l'ontologie différentielle, *via* les relations d'héritage multiple.

Pour que les descriptions produites puissent être effectivement exploitées par le système, nous avons enfin personnellement ajouté aux informations classiques du support de GC formant l'ontologie computationnelle (relation d'héritage, domaines de relations) une série de règles de raisonnement relationnel. Ces règles étaient en effet nécessaires à l'obtention d'un comportement inférentiel satisfaisant par rapport aux besoins de recherche et aux variations informationnelles possibles. Par exemple, on trouvera les règles de transmission de la localisation présentées en page 107, ou bien une règle de « distribution » de la participation à une action : si un *groupe* d'entités est impliqué dans une action, alors on considère que chacun des *membres* de ce groupe est impliqué individuellement dans ladite action.

Au final, l'ontologie de la petite enfance comporte 474 concepts, 68 relations, pour une profondeur hiérarchique maximale de 15 niveaux, ainsi que 106 règles de raisonnement. La figure 5.3 donne un aperçu du contenu de son arborescence conceptuelle.

C'est une démarche semblable qui a été suivie pour le point de vue relatif au traitement documentaire de l'eau dans le domaine géographique. Le corpus retenu comportait une vingtaine de documents, soit 5 heures de vidéos. Des besoins observés, nous avons conçu une ontologie comportant 598 concepts et 55 relations(pour une profondeur hiérarchique maximale de 16) ainsi que 35 règles relationnelles. Nous en donnons un extrait en figure 5.4. Ici, l'accent a été nettement moins mis sur la normalisation du haut niveau ontologique que sur le niveau *structurant* de l'application. Les catégories proposées dans le patron d'indexation (cf. figure 3.13, page 104) correspondent en effet à des entités relevant de domaines relativement différents, et ce à cause du point de vue descriptif retenu. On cherche en effet à faire une description qui concerne à la fois :

- un *thème* donné, en l'occurrence la géographie de l'eau. Ce thème évidemment très vaste comporte des entités de tous types ;
- les objets audiovisuels considérés sous un angle davantage *sémiotique*. Le medium audiovisuel met à disposition des éléments et des procédés permettant de délivrer un message dont certains éléments sont particulièrement intéressants – nos « traits saillants » ;
- une lecture *pédagogique* du contenu audiovisuel : en quoi le message se rapporte au contenu thématique d'une manière plus ou moins appropriée. On insistera à ce niveau sur les valeurs « scientifiques » associées aux différents éléments : définition, explication. . . .

Tous ces éléments que l'on trouve dans notre ontologie, s'ils présentent des points communs au regard du contenu ontologique[8], ont cependant un usage qui les distingue clairement. Plutôt qu'à une seule ontologie homogène, on a affaire à trois sous-ontologies « orthogonales », que les relations du patron d'indexation articulent pour obtenir des descriptions conformes au point de vue retenu. De fait, une de ces branches, celle des procédés et éléments audiovisuels, est d'ailleurs une importation de l'ontologie de l'audiovisuel que nous évoquerons en section 5.2.3.

[8]On trouvera parmi eux des objets spatiaux concrets, des objets temporels, des propriétés, etc.

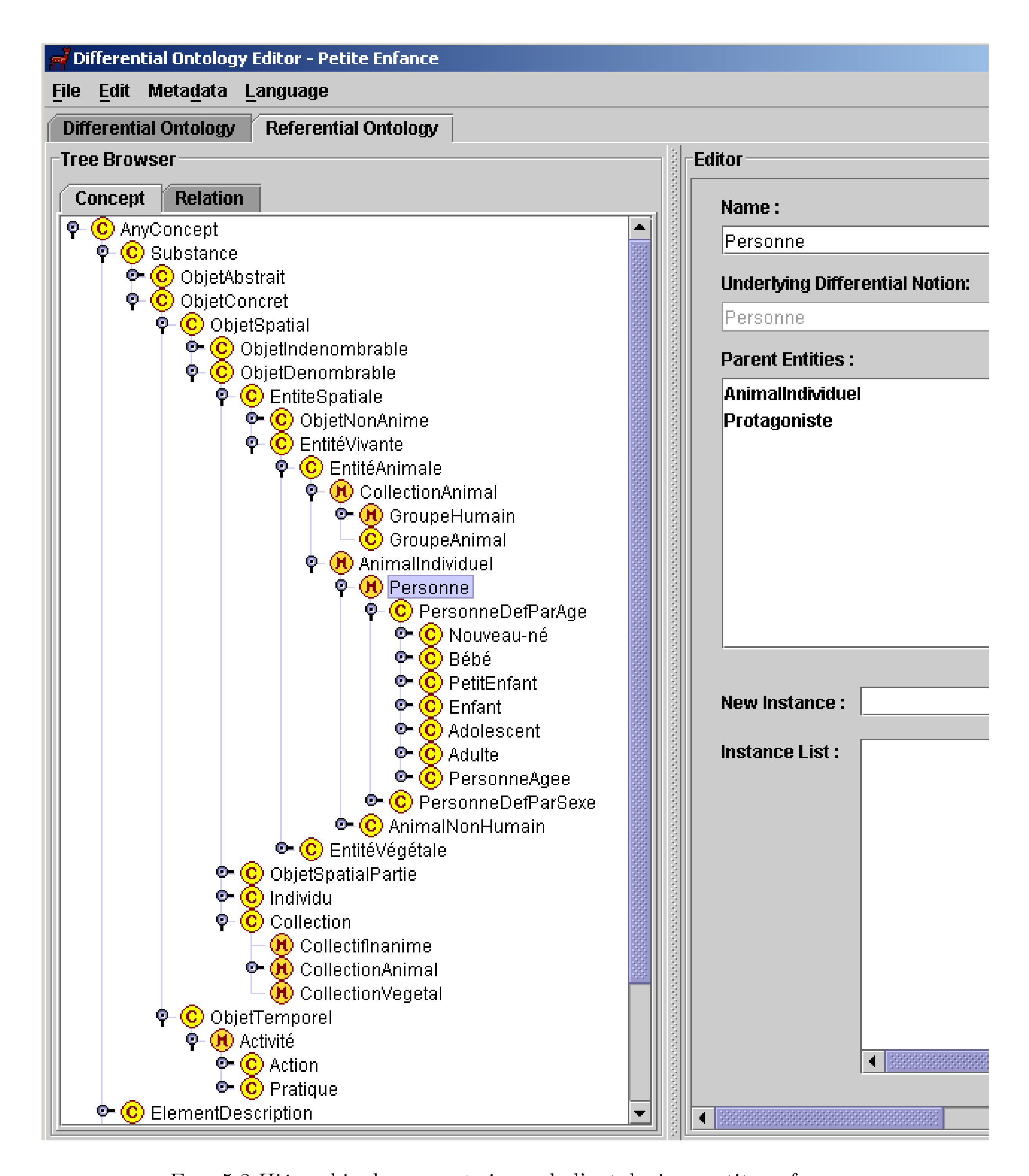


FIG. 5.3 Hiérarchie de concepts issue de l'ontologie « petite enfance »

L'étiquette $\mathcal{M}$ sur un concept indique un cas d'héritage multiple

Il faut noter tout de même que ceci résulte moins de critères de construction ontologique rigoureux que d'un choix caractéristique de conception en fonction de l'application retenue. Il serait en effet tout à fait envisageable de faire *fusionner* les trois branches en une seule, mais nous avons jugé que l'intérêt en termes d'utilisabilité était loin de justifier l'effort requis. D'autant plus que si nous n'avons pas respecté nos principes au plus haut niveau, nous nous sommes efforcé de les appliquer dans les deux branches[9] qui nécessitaient de tels efforts, comme la figure 5.4 le montre partiellement.

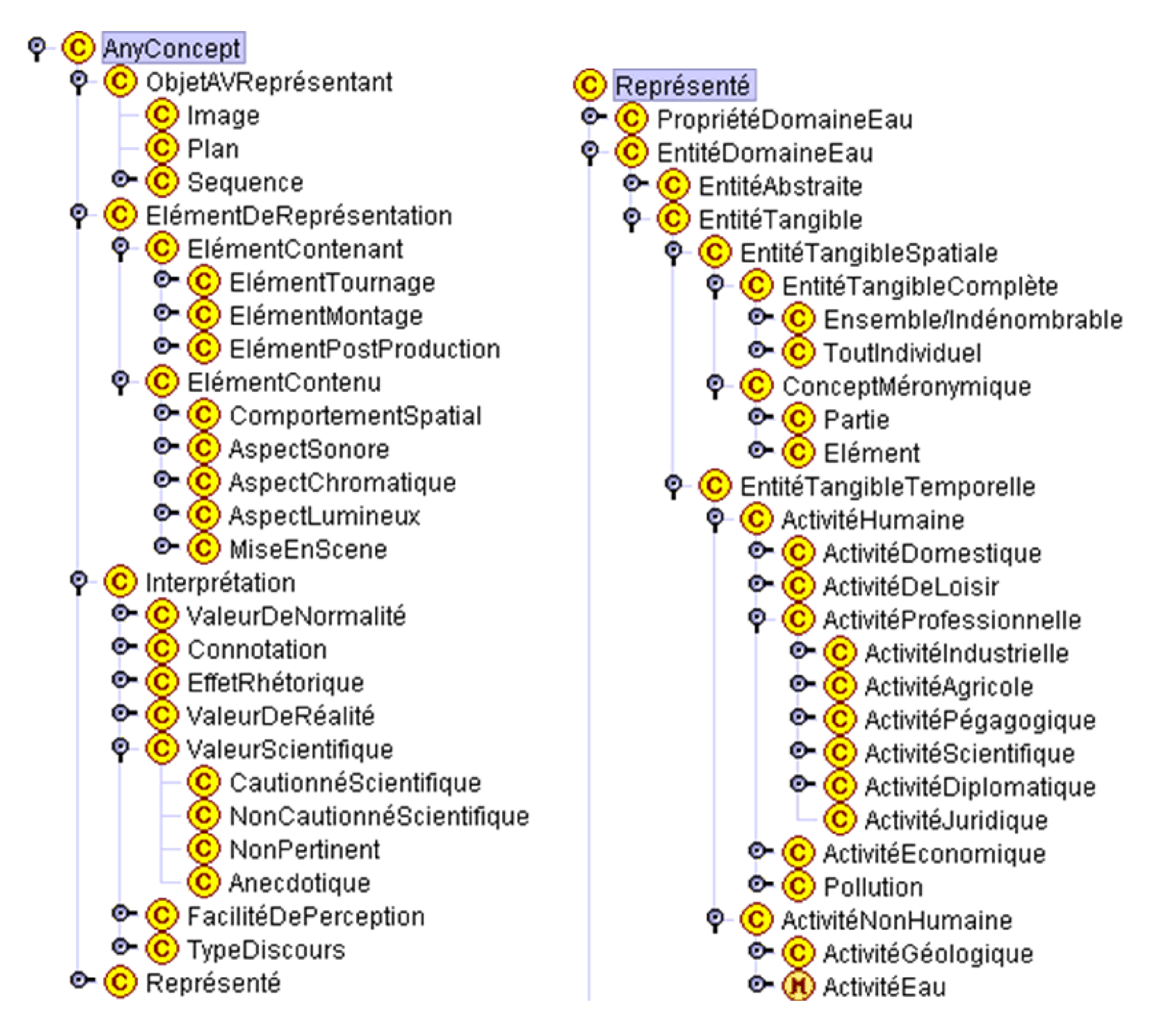


FIG. 5.4 Hiérarchie de concepts issue de l'ontologie de l'« eau »

Après avoir défini les ressources ontologiques du projet, nous avons également participé aux séances d'évaluation de la plate-forme d'indexation d'OPALES[10]. Au cours de ces séances – une

[9] La branche du thème de l'eau proprement dit, et celle des procédés et éléments audiovisuels.

[10] Voir [ICG+04] pour une présentation succincte des interfaces et fonctionnalités, [NNK03] pour une présentation détaillée de l'architecture sous-jacente.

vingtaine d'heures par point de vue – il a fallu familiariser des utilisateurs[11] novices en représentation des connaissances avec les enjeux de ce type d'approche : hiérarchie organisée de façon rigoureuse, index structurés, représentation formalisée exploitable par des mécanismes d'inférence, contrôle du contenu des descriptions... Ces experts des domaines concernés ont ensuite réalisé sous notre supervision une quantité tout à fait réaliste – au regard des volumes documentaires considérés – d'index conceptuels : une trentaine pour le point de vue de la petite enfance, et une quarantaine pour celui de l'eau, ces index comportant en moyenne une dizaine d'entités. Ils ont enfin effectué quelques sessions – informelles – de recherche pour tester les possibilités apportées par les capacités de raisonnement du SBC.

Comme évoqué dans [ICG+04] et détaillé dans [SCC03], ces utilisateurs ont très vite reconnu l'apport d'une méthode de description qui met en valeur les relations entre entités décrites. Les possibilités offertes par les inférences s'appuyant sur des spécifications formelles ont été appréciées : comme on l'a déjà démontré, il devient en effet possible de découpler la forme des requêtes de celle des descriptions, deux contenus similaires sur le plan du fonds pouvant à présent être appariés grâce aux connaissances de raisonnement transférant des informations entre individus des graphes.

Cependant, l'utilisation des ontologies a soulevé de nombreux problèmes dans OPALES. De fait, le constat de complexité que nous avons établi en ce qui concerne l'utilisation des ontologies par des utilisateurs non-experts en RC repose en grande partie sur les observations que nous y avons réalisées. Premièrement, la compréhension de concepts et de relations bien plus précis que des mots-clefs est délicate. A ce titre, les spécifications langagières recommandées par la méthodologie de Bruno Bachimont sont d'une grande aide : on fixe un contexte d'interprétation métier qui simplifie l'accès aux notions. Cependant, l'intérêt de ces définitions différentielles est plus évident en ce qui concerne la phase de conception de l'ontologie. L'organisation suivant des termes précis d'une ontologie s'avère en effet contre-intuitive du point de vue de l'utilisateur non spécialiste en RC, plus habitué aux classifications lâches, davantage guidées par l'usage, présentées dans les thesauri. Si une définition naturelle par la création d'un contexte local pour l'interprétation aide grandement, elle ne peut compenser la complexité apportée par l'utilisation des concepts et des relations abstraits introduits dans l'ontologie pour clarifier la définition rigoureuse des entités métiers. Par exemple, les notions proches d'un point de vue thématique peuvent se retrouver assez éloignées si on recourt à une navigation classique dans l'arborescence différentielle : dans le cas de la petite enfance, les personnes n'étaient pas immédiatement liées aux groupes de personnes – la distinction entre individu et collection étant introduite préalablement – alors que ces deux types d'entités peuvent être considérés au même titre comme les protagonistes d'une action, et sont quasiment interchangeables dans les sommets concernés du patron d'indexation.

Un tel constat ferait donc pencher en faveur de la non-intégration de l'ontologie métier avec une ontologie de haut niveau, comme cela a été le cas pour l'ontologie de l'eau. Ou du moins, puisqu'un tel rattachement est nécessaire à la rigueur de la conceptualisation, veiller à *modulariser* convenablement les ressources, de manière à ne présenter à l'utilisateur les notions les plus abstraites que lorsqu'il en fait la demande explicite.

L'introduction de « points d'entrée » dans l'ontologie référentielle, tels que le concept de `Protagoniste` que nous avons évoqué dans le cas de l'ontologie de la petite enfance, peut aussi constituer un moyen de réconcilier pragmatisme et exigence de cohérence. De tels concepts, généralisant les notions du patron de façon à les considérer d'un point de vue plus fonctionnel,

[11] Ces utilisateurs étaient néanmoins experts dans les domaines retenus : documentalistes du CNRS pour le point de vue de la petite enfance, et professeurs du CNDP pour celui de l'eau.

facilitent la navigation des utilisateurs humains dans des espaces conceptuels davantage déterminés par la proximité d'usage.

Ces rajouts de concepts fonctionnels sont à rapprocher de ce qui a lieu, pour la conception ontologique, lorsqu'on considère des patrons de conception de haut niveau. On avait bien dans cette situation des structures dédiées à un usage, que l'on essayait de rattacher aux concepts et relations d'un domaine donné. Dans un certains sens, nos expérimentations dans OPALES valident bien ce type d'approche : les concepts métiers ont besoin d'un ancrage fonctionnel pour expliciter leur rôle dans l'application, comme ils ont besoin d'un ancrage ontologique *stricto sensu* pour définir leur essence. Mais il ne faut pas oublier, comme on l'a montré dans le chapitre 4, qu'il faut utiliser une structure fonctionnelle qui ne s'écarte pas des pratiques de l'application. *In fine*, c'est la conjugaison des graphes patrons, du contexte d'interprétation global défini par l'ontologie référentielle ainsi présentée et des contextes locaux définis dans l'ontologie différentielle qui a été validée d'un point de vue qualitatif : ainsi guidés, les utilisateurs dans notre projet ont éprouvé beaucoup moins de difficultés à s'approprier la démarche d'indexation conceptuelle qui leur était proposée.

On notera finalement que la souplesse offerte par la combinaison de graphes patrons flexibles et de règles d'inférence a permis de se limiter à un patron par point de vue applicatif, tout en favorisant la création de descriptions très variées. Ceci a facilité de manière importante l'appropriation de l'ensemble de la démarche par nos utilisateurs non-experts.

5.2.3 Expérimentation « chirurgie cardiaque »

Au cours de notre thèse, nous avons pu observer des évolutions techniques majeures, notamment dans le domaine du web sémantique : nouveaux outils, nouveaux langages... Pour être vraiment complète, la recherche qui a été présentée dans ce manuscrit se devait de considérer quelques-unes de ces propositions. Cette expérimentation est venue conclure un long travail de collaboration avec Raphaël Troncy autour de la conception et de l'emploi d'une ontologie de l'audiovisuel qui tirerait parti du savoir-faire documentaire de l'INA dans un cadre formel de description ontologique et de recherche automatisée d'index [IT04, IT05b][12].

Mise en place de l'expérimentation

Pour demeurer dans un cadre familier, nous avons choisi d'adapter un thème déjà abordé, celui d'une lecture « critique » des documents audiovisuels, telle qu'abordée dans le point de vue « eau » d'OPALES. Cette préoccupation est relativement proche des préoccupations documentaires de notre Institut, puisqu'elle implique une description précise de la structure et du contenu thématique des documents audiovisuels. De ce fait, elle intéressait aussi au premier plan les travaux effectués au cours de la thèse de Raphaël Troncy. Dans ces travaux était en effet particulièrement reconnu le fait que les descriptions doivent mélanger des éléments strictement orientés « description audiovisuelle » et des notions relevant d'un domaine donné.

Nous souhaitions cependant changer de domaine. Nous avons fait le choix de nous concentrer sur le celui de la médecine. Des fonds de l'INA, nous avons extrait une trentaine de documents AV, en majorité des magazines, d'à peu près une heure chacun. Nous en avons ensuite sélectionné la moitié liée aux thèmes spécifiques du cœur ou de la chirurgie cardiaque. Il s'agit donc d'une collection plutôt homogène, ce qui facilite la recherche d'une ontologie thématique adaptée. De fait, comme le sujet de la médecine a attiré l'attention d'un grand nombre de chercheurs ces dernières années, une grande quantité de ressources ontologiques sont disponibles dans ce domaine.

[12] Pour une version francisée de [IT05b], on peut se reporter à [IT05a].

De plus, comme ces programmes devaient être diffusés sur des chaînes généralistes, ce sont aussi de bons exemples sur la façon dont les procédés audiovisuels sont utilisés pour vulgariser des sujets scientifiques complexes.

Ressources ontologiques

La première des ontologies que nous avons employées ici est celle de l'audiovisuel. Initiée par Raphaël Troncy, cette ontologie est le fruit d'un certain nombre de mois d'efforts communs pour formaliser les pratiques en cours dans notre domaine très large. Elle a été conçue selon les points méthodologiques que nous avons évoqués dans cette thèse : sa normalisation, sa formalisation et son opérationnalisation ont été effectuées suivant les principes de Bruno Bachimont tels que repris dans notre éditeur **DOE**, et elle reprend, en les adaptant, les patrons de conception de Gangemi et ses collègues, tels que nous les avons présentés dans le chapitre 4 de cette thèse.

Cette ontologie se concentre d'abord sur la caractérisation d'éléments documentaires : le concept principal est celui d'*objet de production* AV, qui représente la notion même de document AV. La première distinction intervient entre *programmes* (entités plutôt indépendantes du point de vue de la production et de la diffusion) et *séquences* (parties de programmes ou d'autres séquences). Ces concepts sont ensuite spécialisés en fonction de traits différentiels liés à la forme ou au contenu : on obtient ainsi une hiérarchie de *genres* audiovisuels. Par exemple, les programmes sont répartis entre *composites* et *simples*[13], les premiers, contrairement aux seconds, étant composés d'une suite d'éléments autonomes du point de vue de la forme et du contenu. Ils sont ensuite classés suivant leur longueur et leur contenu général (fiction, information, divertissement). Après quelques étapes additionnelles de spécialisation, on peut trouver les genres télévisuels courants : *comédie de situation*, *documentaire*, *spectacle TV*...

L'ontologie introduit également les notions utilisées pour préciser les caractéristiques des objets AV. Tout d'abord, nous avons introduit une hiérarchie des rôles que les personnes peuvent jouer dans un programme, soit en tant qu'auteurs (*producteur*, *réalisateur*) mentionnés à cause de leur importance dans la production du programme, soit comme participants (*animateur*, *acteur*), apparaissant dans la description parce que visibles dans le document. Ensuite, nous pouvons trouver un ensemble important de propriétés qui reflètent diverses préoccupations ou modalités de la production (filmage, comme pour *mouvement de caméra* ; montage ou post-production, comme pour *insertion de texte*) et la diffusion (*date de diffusion*, *périodicité*, *public visé*...).

Le fait que la conception de l'ontologie ait répondu à des critères méthodologiques précis permet de l'étendre assez facilement pour l'adapter aux besoins applicatifs à venir. Et si, pour les besoins d'OPALES, nous en avions intégré une première version dans celle d'un de nos points de vue, lors de cette expérimentation nous avons réalisé l'opération inverse, en ajoutant des relations basiques qui dénotent des jugements interprétatifs concernant la manière dont le contenu thématique est présenté par ces documents : *clarifie*, *exemplifie*, *démontre*... Finalement, nous proposons des relations conceptuelles liant les objets AV à des thèmes externes, ce qui permet la description du contenu proprement dit.

Pour cette description du contenu thématique, nous nous sommes rapidement tourné vers l'ontologie MENELAS, qui décrit le domaine des pathologies coronariennes [ZC94], et contient une importante quantité de concepts liés à la chirurgie cardiaque. Ceci nous permettait de réutiliser une ontologie déjà accessible – celle de MENELAS – tout en appliquant notre démarche à des notions que nous n'avions pas déjà utilisées : les précédentes expérimentations n'avaient en effet repris de cette ontologie que son *haut niveau*, sans se préoccuper de ses notions plus spécialisées.

[13]Un extrait de la hiérarchie de ces programmes est représenté dans la figure 2.7 de la page 56.

Se pose ensuite la question du traitement effectué sur ces ontologies avant de procéder à l'indexation. Faut-il, comme dans OPALES, créer un référentiel conceptuel unifié, permettant de rendre compte de manière complète mais rigide d'un point de vue descriptif ? De fait, plutôt que de *fusionner* (cf. page 114) les deux ontologies utilisées ici, nous avons choisi de les garder distinctes, et introduit quelques équivalences de classes quand cela était nécessaire. Par exemple, on peut avec les langages du web sémantique que nous avons utilisés (on verra le détail par la suite) affirmer que `av :person` et `menelas :human_being` sont des concepts équivalents, en tout cas du point de vue extensionnel – les ensembles d'individus qu'ils dénotent sont égaux. Ceci permet d'articuler l'ontologie de l'audiovisuel et l'ontologie thématique sans que les conceptualisations respectives dont celles-ci relèvent soient rendues plus confuses par une trop grande proximité.

Il est intéressant de noter que cette possibilité de faire référence explicitement à plusieurs ontologies pour produire les descriptions est un point qui distingue l'expérimentation proposée ici de celle menée dans OPALES, même si pour l'un des points de vues nous nous étions déjà refusé à faire une fusion complète[14]. A ce titre, cette expérimentation constitue bien une évolution vers l'utilisation de langages et d'outils liés au web sémantique, profitant ainsi des nombreux efforts de recherche menés dans ce domaine.

Indexation des documents AV

Dans notre point de vue, décrire un document AV implique de considérer des aspects documentaires – identifier les éléments – et thématiques – affirmer que ces éléments sont *à propos* de quelque chose. Distinguer l'ontologie AV des autres ontologies thématiques nous permet de considérer ces deux aspects.

L'outil **SegmenTool**[15], illustré en figure 5.5, nous a permis de segmenter les documents AV et de créer les méta-données spécifiques à la description documentaire audiovisuelle[16], en utilisant le langage MPEG-7 [MPEG-701].

Pour faciliter le processus de description et rendre ses résultats plus cohérents, nous avons évidemment utilisé un patron d'indexation. Dans le cadre de notre expérimentation, nous nous sommes encore une fois appuyés sur le point de vue « eau » du projet OPALES, ainsi que sur les besoins d'analyse documentaire de l'INA dont une formalisation avait été proposée dans [IT04]. Comme évoqué dans la figure 5.6, il faut décrire les éléments AV en leur assignant certaines valeurs de propriétés (par exemple, la manière dont ils sont produits) et en les décomposant d'un point de vue documentaire. Leur contenu doit également être indexé par l'assertion de relations avec des concepts du domaine, relations de nature strictement représentationnelle – ce qui est montré dans les vidéos – ou plus interprétatives – quelle est l'utilisation de ces représentations.

Comme on l'a déjà vu, un tel patron peut en fait engendrer des descriptions extrêmement riches, comme celle représentée par le code 5.1. Nous préférons d'ailleurs présenter ici – en figure 5.7 – une version « graphique » et quelque peu simplifiée de cet index[17] qui permet de mettre en valeur la distinction entre les deux types de connaissances qui nous intéressaient dans

[14] Le fait est que pour les besoins du formalisme de représentation retenu il nous a bien fallu intégrer ces différentes branches au sein d'une même ontologie, ce qui, sans pour autant gêner son utilisation dans le point de vue applicatif retenu, a pu nuire à la cohérence théorique du résultat final.

[15] Cet outil a été développé par l'équipe DCA de la direction de la recherche de l'INA et a été partiellement financé dans le cadre du projet PRIAMM CHAPERON.

[16] Raphaël Troncy propose en effet d'ancrer les descriptions conceptuelles dans un langage strictement documentaire [Tro04].

[17] On remarquera au passage que, pure coïncidence, nous retrouvons là un exemple extrêmement proche de celui que nous avions utilisé au chapitre 1, si l'on fait abstraction des intitulés anglais des noms des concepts et des relations.

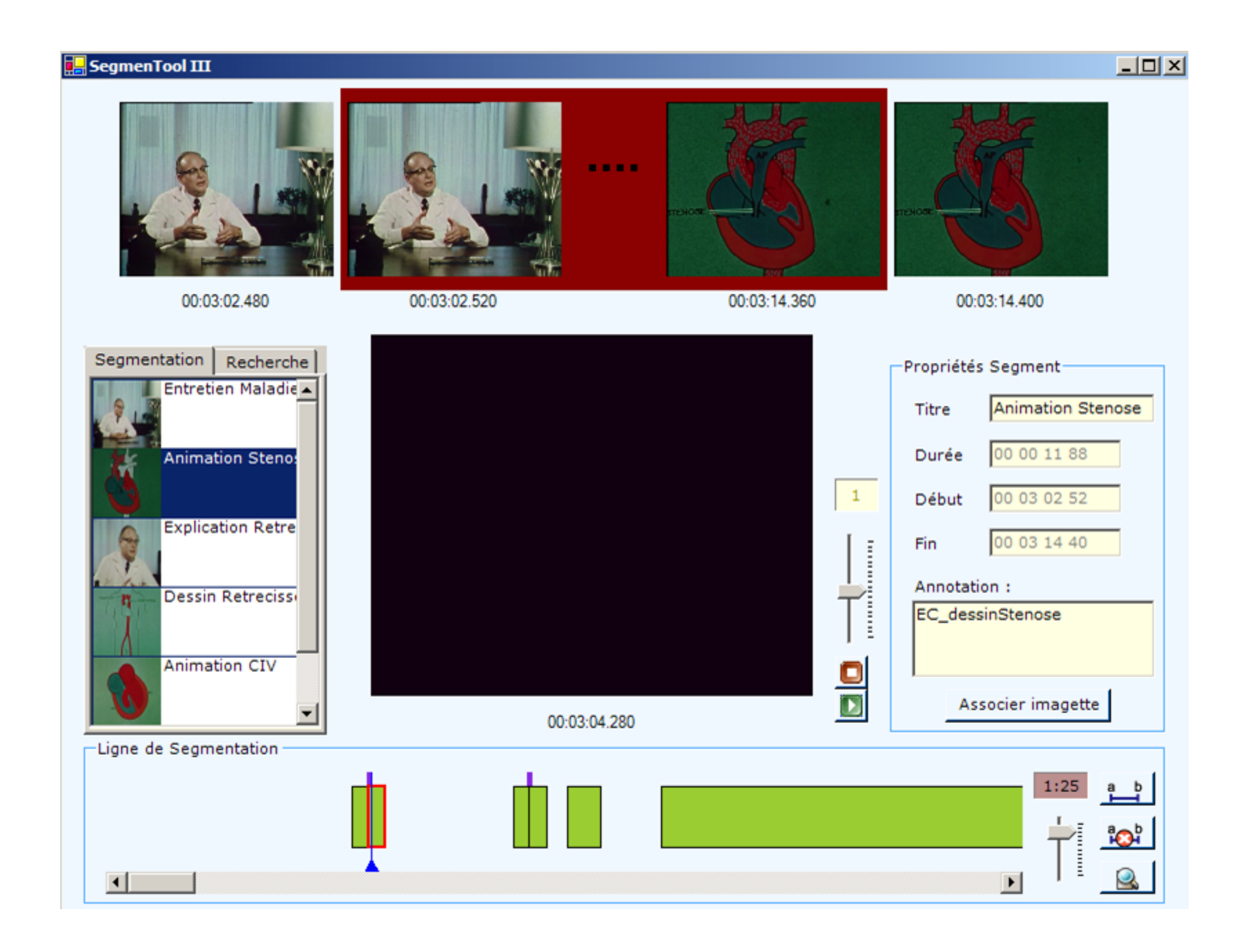


FIG. 5.5 Utilisation de l'outil **SegmentTool** pour produire une structure documentaire

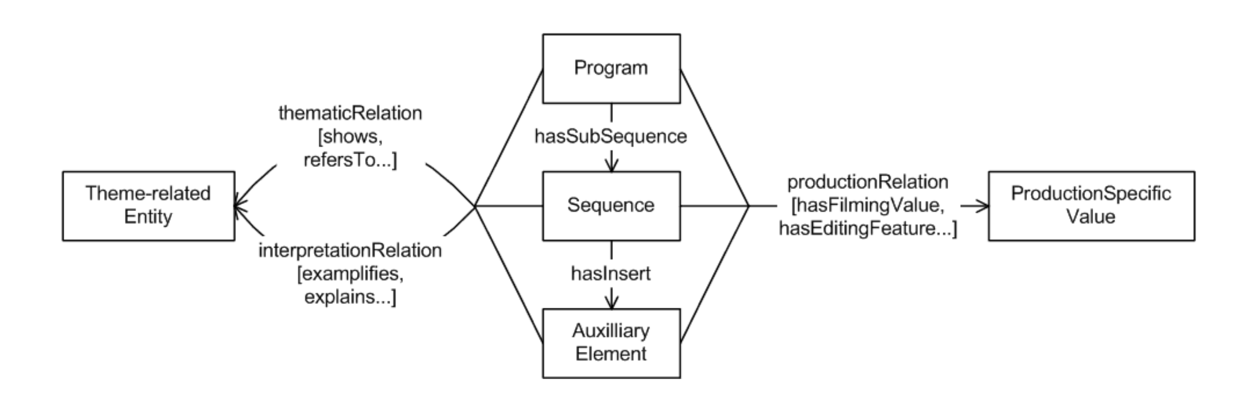


FIG. 5.6 Patron de description relationnel de l'expérimentation « chirurgie cardiaque »

cette expérience, à savoir celle liée strictement à l'audiovisuel (**AV**) et celles liées à la médecine cardiaque (**Thème**).

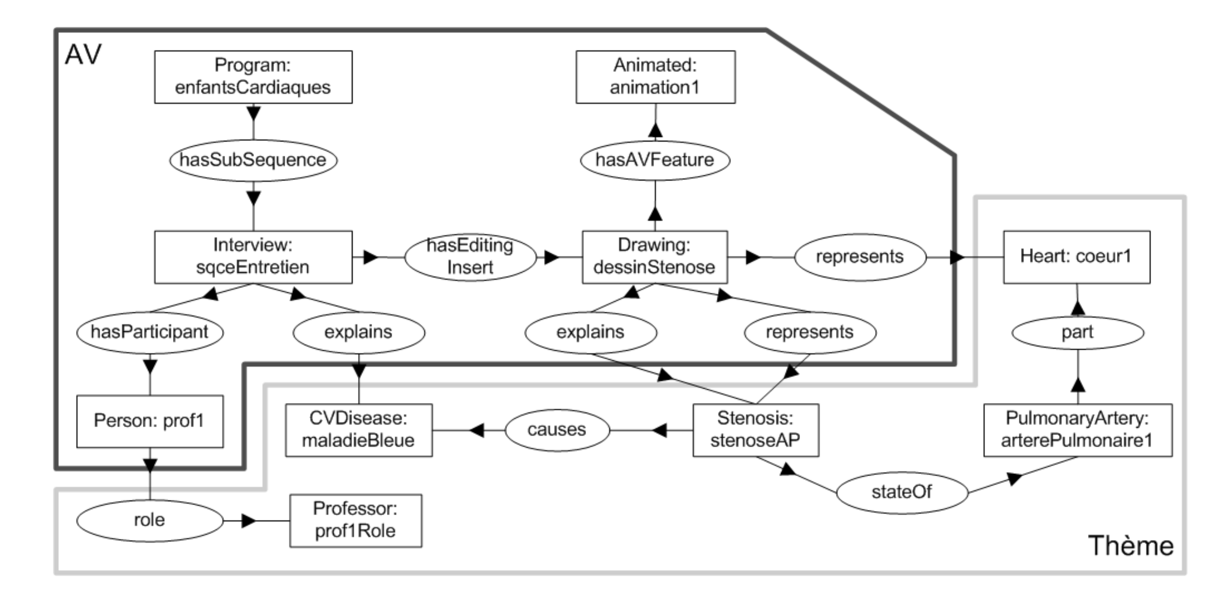


Fig. 5.7 Représentation graphique de l'index du code 5.1

Rechercher et raisonner dans la base de connaissances

Comme on l'a vu au fil de ce manuscrit, l'intérêt de l'utilisation des ontologies dans un système d'indexation et de recherche repose en grande partie sur l'utilisation de processus d'inférence exploitant les connaissances de l'ontologie pour améliorer l'accès aux index.

Les assertions explicites que l'on trouve dans les descriptions peuvent en effet être complétées par des assertions *dérivées*. Ainsi, il serait souhaitable, à partir de l'index de la figure 5.7, d'obtenir dans la base de connaissances des assertions factuelles correspondant à ce qui est représenté en figure 5.8. Par exemple, si une séquence contient une séquence expliquant un sujet donné, on peut en déduire qu'elle aussi contribue à l'explication de ce sujet. On peut ainsi retrouver des objets AV qui font référence à des thèmes variés, même si lesdits thèmes n'apparaissaient explicitement que dans les éléments contenus dans ces objets. Ici, le système jugera notre documentaire pertinent pour une requête telle que « trouver un programme qui explique une maladie et offre une représentation visuelle d'une de ses causes ».

Quelles sont donc les possibilités offertes en la matière par le web sémantique ? Quels raisonnements pouvons-nous rendre possibles, en fonction du langage de représentation des index, mais aussi de celui de spécification d'ontologies ? Nous ne pouvions pas évidemment tester toutes les solutions offertes. Par contre, en présence d'un mouvement certain de standardisation des langages de représentation de connaissances sur le web, encouragé par le W3C[18], nous avons trouvé pertinent de nous attacher aux propositions les plus matures, RDF pour la représentation des faits relationnels qui constituent les index, puis RDFS et OWL pour la représentation des connaissances ontologiques.

[18] http://www.w3c.org

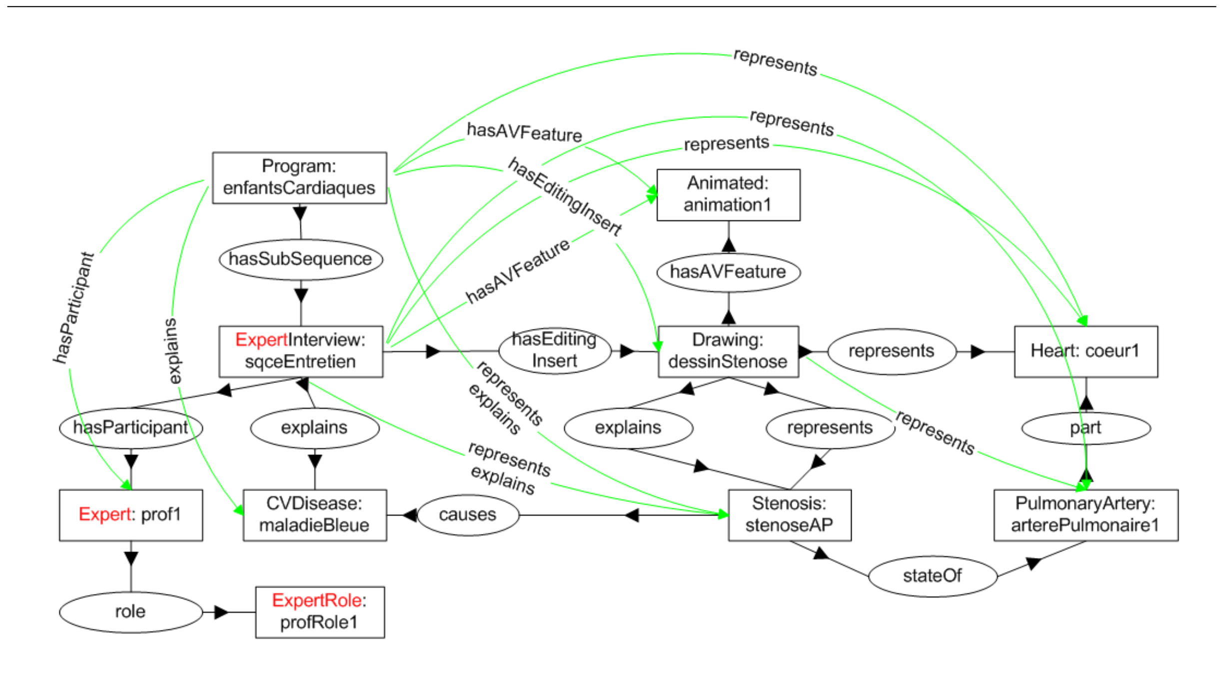


Fig. 5.8 Index complété avec des connaissances inférées

En première approche[19], il faut retenir que RDF est un langage de représentation d'essence relationnelle : les faits y sont représentés sous forme de triplets *sujet–propriété–objet*. RDF propose un petit nombre de propriétés standards qui permettent de définir[20] les entités considérées. Par exemple, on pourra donner à un individu un *type* en utilisant la propriété `rdf :type` – ou, comme dans le code 5.1, une balise marquée par le nom du type. RDF, en tant que langage de représentation de connaissances orienté vers l'utilisation des réseaux, permet le recours à toutes les fonctionnalités offertes par la syntaxe XML [XML04] sur laquelle il repose. Il devient en particulier possible de faire référence à des objets situés dans des *espaces de nommage* distants, ce qui permet entre autres d'utiliser plusieurs ontologies ou bases de faits physiquement déconnectées dans une seule description. Le code 5.1 donne la représentation de notre index favori dans ce langage.

RDF, par son emploi de types et de propriétés, appelle évidemment la création des référentiels permettant justement d'introduire de tels objets en tant que langage de description propre à une préoccupation particulière. Pour cela, le langage de représentation de schémas[21] RDFS – pour RDF-SCHEMA – a été créé. Nous ne rentrerons pas dans les détails de cette norme – le lecteur peut se reporter à [RDF04b] – mais on peut toutefois en signaler les fonctionnalités principales du point de vue de la représentation des connaissances :

- définition de classes (instances de la primitive `rdfs:Class`) qui permettront de typer les objets d'une description ;
- organisation de ces classes par des liens hiérarchiques (`rdfs:subClassOf`) à la sémantique

[19]Pour une présentation complète de RDF (*Resource Description Framework*), on pourra se reporter à [RDF04a].

[20]Bien que d'un point de vue de la représentation fondamentale rien ne distingue ces « définitions », correspondant elles aussi à des triplets, des assertions « classiques ».

[21]Dans notre terminologie, il s'agit donc d'un *méta*-langage de représentation, puisqu'il définit les primitives de connaissances qui seront utilisées ensuite pour représenter les faits d'une application donnée. Ici, le terme *schéma* est à comprendre au sens des schémas XML, ces spécifications – syntaxiques, puisque XML n'a pas vocation à représenter le sens des objets qu'il permet de manipuler – qui définissent des règles permettant de valider des fichiers XML.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:av="http://www.ina.fr/Audiovisual/"
    xmlns:menelas="http://www.ina.fr/Menelas/">

    <av:Program rdf:about="JNV_enfantsCardiaques">
        <av:hasSubSequence rdf:resource="EC_sqceEntretienMaladiesCV"/>
    </av:Program>

    <av:Interview rdf:about="EC_sqceEntretienMaladiesCV">
        <av:hasParticipant rdf:resource="EC_prof1"/>
        <av:hasEditingInsert rdf:resource="EC_dessinStenose"/>
    </av:Interview>

    <av:Person rdf:about="EC_prof1">
        <menelas:role rdf:resource="EC_profRole1"/>
    </av:Person>
    <menelas:professor rdf:about="EC_profRole1"/>

    <menelas:cardio_vascular_disease rdf:about="maladieBleue"/>

    <av:Explanation rdf:about="EC_explicationMaladiesCV">
        <av:explains rdf:resource="maladieBleue"/>
    </av:Explanation>

    <av:Drawing rdf:about="EC_dessinStenose">
        <av:explains rdf:resource="EC_stenoseAP"/>
        <av:hasAVFeature rdf:resource="animatedValue"/>
        <av:represents rdf:resource="EC_stenose"/>
        <av:represents rdf:resource="EC_coeur1"/>
    </av:Drawing>

    <menelas:stenosis rdf:about="EC_stenoseAP">
        <menelas:cause_of rdf:resource="maladieBleue"/>
        <menelas:state_of rdf:resource="EC_arterePulmonaire1"/>
    </menelas:stenosis>

    <menelas:pulmonary_artery rdf:about="EC_arterePulmonaire1">
        <menelas:part rdf:resource="EC_coeur1"/>
    </menelas:pulmonary_artery>

    <menelas:heart rdf:about="EC_coeur1"/>

    <av:Animated rdf:about="EC_animation1"/>

</rdf:RDF>
```

Code 5.1 – Index décrivant une émission contenant une interview, en RDF

précise ;
– organisation des propriétés par un lien hiérarchique similaire (`rdfs:subPropertyOf`) ;
– assignation de domaines et co-domaines aux propriétés (`rdfs:domain` et `rdfs:range`), permettant de restreindre leur application à des classes d'objets particulières.

De fait, les fonctionnalités de RDFS le placent à un niveau d'expressivité comparable à celui des supports simples de graphes conceptuels que nous avons présentés en section 2.2.2. Des travaux comme [Bag03] essaient d'ailleurs de rendre compte des similarités existantes, tant du point de vue des descriptions autorisées que des inférences effectuées. Le code 5.2 donne un bref exemple d'utilisation de RDFS pour notre expérimentation, exemple qui définit une classe et une propriété que nous connaissons déjà bien. . .

```
...
<rdfs:Class rdf:about="Interview">
    <rdfs:label xml:lang="en">Interview</rdfs:label>
    <rdfs:subClassOf rdf:resource="DialogSequence"/>
</rdfs:Class>
...
<rdf:Property rdf:about="hasParticipant">
    <rdfs:label xml:lang="en">hasParticipant</rdfs:label>
    <rdfs:subPropertyOf rdf:resource="AVObjectXPerson_Conditions"/>
    <rdfs:domain rdf:resource="AVObject"/>
    <rdfs:range rdf:resource="Person"/>
</rdf:Property>
...
```

CODE 5.2 – Définitions de la classe `Interview` et de la propriété `hasParticipant` en RDFS

Pour stocker et interroger les ontologies et les assertions, nous utilisons l'architecture **Sesame** [BKv02]. Pour l'instant, celle-ci utilise RDFS comme langage de représentation d'ontologies, et offre des services de raisonnement conformes aux spécifications de la théorie des modèles RDF. Ceci autorise des inférences basiques, comme l'utilisation des liens de subsomption pour les concepts et les relations. Dans notre exemple, il est ainsi possible de trouver l'interview décrite dans les résultats d'une recherche de « séquences expliquant la maladie bleue », puisque dans l'ontologie de l'AV `Interview` spécialise `Sequence`.

Cependant, cela n'est pas suffisant pour l'exploitation que nous désirons, précise, et utilisant à la fois les propriétés des concepts et des relations présents dans les index. Nous nous plaçons en effet toujours dans un cadre comme celui d'OPALES, où nous avions jugé pertinent d'utiliser en plus des fonctionnalités des supports basiques des graphes conceptuels celles de règles de raisonnement plus complexes. Dans le contexte du web sémantique, l'opportunité d'utiliser les possibilités des langages OWL[22] – ou au moins celles du sous-ensemble décidable OWL-DL – semble en particulier séduisante. Avec OWL, on peut préciser qu'une `ExpertInterview` est exactement définie comme une interview où *au moins un* participant joue un rôle d'expert. Le concept `ExpertRole` sera lui défini par le biais d'une équivalence de classes énumérant les rôles du domaine qui peuvent être considérés comme dénotant une certaine expertise pour l'application. Dans notre cas, nous avons sélectionné parmi les spécialisations du concept `role` de l'ontologie

[22] Voir en page 60, ou [OWL04] pour une présentation complète.

MENELAS les concepts spécialisant les rôles *académique*, *professionnel* et *hospitalier*, à l'exception du rôle attribué à l'institution hospitalière elle-même, ce qui a été rendu possible en utilisant le constructeur OWL `owl:complementOf`. Cet exemple, formulé partiellement en logique de description dans la figure 2.12 de la page 61, est présenté en utilisant la syntaxe XML de OWL dans le code 5.3. Ainsi, on peut encoder les connaissances autorisant les inférences illustrées par les changement de types de concept sur la figure 5.8, et répondre à des requêtes comme « émissions contenant des témoignages d'expert, expliquant une maladie cardio-vasculaire ». Pour implémenter de tels raisonnements, on peut se tourner vers des raisonneurs OWL comme **BOR**[23] [SJ02], qui a été intégré dans Sesame.

Et pourtant, cette solution peut ne pas être assez puissante. Comme nous ciblons des index riches en connaissances relationnelles, nous voudrions exploiter des connaissances de raisonnement exploitant ces relations. OWL-DL permet de spécifier des propriétés algébriques pour les relations, comme la transitivité : le code 5.4 montre par exemple comment on peut spécifier que la propriété `hasSubSequence` est transitive, spécification qui pourra être utilisé par un raisonneur comme **BOR**. Ceci est clairement utile, mais on manque de possibilités pour encoder des connaissances plus générales, relatives en particulier à la composition des relations. Nous avons par exemple besoin de créer des règles représentant des formules comme[24]

$\forall x, y, z$ `hasSubSequence(x,y)` $\wedge$ `represents(y,z)` $\Rightarrow$ `represents(x,z)`.
On pourrait ainsi traiter une requête semblable à la précédente, mais demandant en sus que l'émission soit « illustrée par des images montrant l'objet concerné par la pathologie ».

Ces préoccupations sont reconnues dans la communauté du web sémantique, et commencent à se voir traitées par des langages et des outils appropriés. Ainsi, SWRL [HPB+04] constitue un pas vers le rapprochement entre les langages OWL et les règles logiques. Quelques-unes des ces propositions peuvent être implémentées *via* des cadres logiques décidables comme OWL-DLP [GHVD03] qui restreint OWL-DL tout en autorisant le recours à certains des éléments des programmes logiques, au nombre desquels figurent les règles de raisonnement que nous avons qualifiées de relationnelles. Perdre une partie de l'expressivité OWL – en particulier les restrictions existentielles dans les conditions nécessaires – peut se révéler gênant, mais la richesse des règles relationnelles possibles[25] nous fait préférer un tel choix.

Dans **Sesame**, une telle opportunité est implémentée dans un module d'inférence *sur mesure*, où les axiomes et les règles de RDFS sont complétées par ceux de OWL-DLP et des règles de raisonnement spécifiques aux ontologies exploitées. C'est une manière plutôt rigide de concevoir un tel système – les règles sont encodées au niveau de la spécification du raisonneur, et non dans l'ontologie elle-même – mais elle permet cependant de mettre en œuvre des raisonnements intéressants. On peut ainsi, de l'index de la figure 5.8, déduire les assertions grisées et en pointillé qui viennent s'ajouter à celles issues des raisonnements OWL. On dispose alors de bien plus d'éléments dans la base de connaissances pour répondre aux requêtes. Le tableau 5.2 résume le nombre de triplets (explicites et inférés) contenus dans la base de connaissances Sesame pour notre expérimentation. La saturation de cette base est obtenue en utilisant les règles OWL-DLP complétées par une vingtaine de règles spécifiques à l'application, en majorité des règles de composition. Le code de la figure 5.5 montre un extrait du fichier spécifiant le fonctionnement du raisonneur sur mesure que nous avons construit pour **Sesame**. Cette règle s'intéresse à la composition des relations de décomposition en séquence et de représentation : si un objet a comme

[23] **BOR** implémente en fait la sémantique du langage DAML+OIL, mais celle-ci est extrêmement proche de ce qui peut être spécifié pour les moteurs OWL.
[24] Pour d'autres exemples – simplifiés – on peut se reporter à la page 65.
[25] Ces règles peuvent d'ailleurs remplacer partiellement les éléments perdus de OWL-DL, comme dans le cas des restrictions existentielles dans des définitions par condition suffisante.

```
<owl:Class rdf:about="ExpertInterview">
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <owl:Class rdf:about="Interview"/>
        <owl:Restriction>
          <owl:onProperty rdf:resource="hasParticipant"/>
          <owl:someValuesFrom>
             <owl:Class>
                <owl:intersectionOf rdf:parseType="Collection">
                   <owl:Class rdf:about="Person"/>
                   <owl:Restriction>
                      <owl:onProperty rdf:resource="&menelas;role"/>
                      <owl:someValuesFrom rdf:resource="ExpertRole"/>
                   </owl:Restriction>
                </owl:intersectionOf>
             </owl:Class>
          </owl:someValuesFrom>
        </owl:Restriction>
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</owl:Class>
<owl:Class rdf:about="ExpertRole">
  <owl:equivalentClass>
    <owl:Class>
       <owl:intersectionOf rdf:parseType="Collection">
          <owl:Class>
             <owl:unionOf rdf:parseType="Collection">
                <owl:Class rdf:about="&menelas;academic_role_function"/>
                <owl:Class rdf:about="&menelas;hospital_role"/>
                <owl:Class
                   rdf:about="&menelas;professional_role_function"/>
             </owl:unionOf>
          </owl:Class>
          <owl:Class>
             <owl:complementOf>
                <owl:Class rdf:about="&menelas;the_hospital_role"/>
             </owl:complementOf>
          </owl:Class>
       </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</owl:Class>
```

Code 5.3 – Définitions OWL des classes ExpertInterview et ExpertRole

```
<owl:TransitiveProperty rdf:about="hasSubSequence"/>
```

CODE 5.4 – Déclaration de la transitivité de la propriété `hasSubSequence`

sous-séquence un objet qui représente un sujet quelconque, alors le premier objet représente également ce sujet.

	Triplets explicites	**Triplets inférés**	**Total**
Modèle RDF			129
Ontologie AV	5231	10810	16041
Ontologie MENELAS	10534	26637	37171
Instances	276	1507	1783
Total	16041	38954	54995

TAB. 5.2 Nombre de triplets (explicites and inférés) dans la base de connaissances *Sesame*. Le modèle RDF désigne les triplets définissant le langage de représentation lui-même.

Il est évident que l'enjeu de l'exploitation du raisonnement dans un système d'information va au-delà d'un gain mesuré quantitativement, puisqu'il s'agit de simuler, grâce à l'implémentation de connaissances de raisonnement propres à un domaine d'application, une partie des raisonnements[26] qui sont effectués par les recherchistes dans les systèmes documentaires actuels. Nous considérons cependant que, sur une expérimentation assez réduite, nous avons là, en complément de l'index complété de la figure 5.8, un indice assez révélateur de l'intérêt des mécanismes de raisonnement standardisés par l'initiative du web sémantique pour notre approche.

Il faut remarquer que **Sesame** offre la possibilité d'exprimer des règles utilisant la pleine puissance de RDFS, qui considère que l'on peut typer, en plus des individus d'une base de connaissance, les types que l'on assigne à ces objets. Toute entité, quel que soit le niveau d'abstraction dont elle relève, peut toujours être considérée comme un sujet et être à son tour typée, ce qui place de fait RDFS – et **Sesame** – en dehors du cadre de la logique du premier ordre[27]. On peut donc entre autres créer des règles qui sont capables de s'appliquer à des propriétés dont on sait seulement qu'elles spécialisent une propriété donnée. La règle du code 5.5 peut ainsi être formulée comme dans la figure 5.6 : si un objet audiovisuel a comme sous-séquence un autre objet qui représente un sujet quelconque d'une façon bien déterminée, alors le premier objet représente lui aussi ce sujet, et ce, de la même façon. Ceci est important, car on peut alors ne pas réécrire la règle pour toutes les spécialisations de `represents` : en l'état actuel des choses, en effet, la règle ne peut que déduire une information utilisant cette propriété, même si on était face à une propriété plus précise, possiblement inconnue au moment de la définition des règles – une propriété faisant référence à une représentation *symbolique*, par exemple[28]. Mais pour cette expérimentation nous n'avons pas franchi ce pas[29], pour rester dans les limites habituellement respectées dans les

[26]Ces « raisonnements » consistent essentiellement, on le rappelle, en des reformulations de requêtes, pour élargir ou préciser leurs résultats.

[27]Il faut remarquer que les langages les plus simples de la famille OWL, dont OWL-DL, opèrent une restriction des capacités expressives de RDFS, justement pour rester dans le cadre décidable des logiques de description qui relèvent de sous-ensembles de la logique du premier ordre.

[28]Cf. la figure 2.8 de la page 56.

[29]Pas plus que nous n'avons spécifié les règles pour toutes les spécialisations de relations, ce que nous avions fait dans OPALES. Ce problème se pose en effet également dans le cadre des graphes conceptuels, qui respectent

```
<!-- composition_hasSubSequence_represents -->
<rule name="composition_hasSubSequence_represents">
    <premise>
        <subject   var="xxx"/>
        <predicate uri="&av;hasSubSequence"/>
        <object    var="yyy"/>
    </premise>
    <premise>
        <subject   var="yyy"/>
        <predicate uri="&av;represents"/>
        <object    var="zzz"/>
    </premise>
    <consequent>
        <subject   var="xxx"/>
        <predicate uri="&av;represents"/>
        <object    var="zzz"/>
    </consequent>
</rule>
```

Code 5.5 – Déclaration de la composition des propriétés `hasSubsequence` et `represents` dans **Sesame**

travaux du web sémantique.

5.2.4 Expérimentation EON

L'expérimentation suivante, présentée dans l'article [ITM03], a eu lieu dans le cadre de l'évaluation des éditeurs d'ontologies menée pour l'édition 2003 de l'atelier *Evaluation of Ontology Tools*[30]. Cette année-là, l'enjeu était de tester la capacité des éditeurs conçus par la communauté de l'ingénierie ontologique en ce qui concerne l'échange d'ontologies. Il s'agissait plus précisément de voir comment ces éditeurs pouvaient échanger les conceptualisations qu'ils avaient permis de représenter. Pour cela, l'unique moyen est d'exporter et d'importer ces spécifications au moyen des langages standards de représentation d'ontologies. La question qui se pose est alors de savoir comment les informations qui permettent de spécifier les ontologies dans un éditeur peuvent être encodées dans ces langages, et importées par les autres environnements de conception. Ces informations sont évidemment différentes d'un éditeur à l'autre, puisque ceux-ci ont été souvent conçus avec des préoccupations qui, si elles se recoupent, restent distinctes… Nous présenterons en section 5.2.4 les résultats concernant les performances de **DOE** concernant l'interopérabilité, pour nous concentrer maintenant sur l'expérience de modélisation qui a été conduite pour les besoins de cette expérimentation.

la logique du premier ordre, et où une relation ne peut être considérée comme une variable.

[30]Les ateliers EON ont été organisés chaque année depuis 2002, lors de conférences internationales dédiées au web sémantique. Les thèmes privilégiés sont ceux de l'interopérabilité entre éditeurs d'ontologies, les possibilités d'alignement entre ontologies similaires, etc. Les actes de l'atelier 2003, qui se tenait dans le cadre de la conférence ISWC, sont disponibles à l'adresse `http://CEUR-WS.org/Vol-87/`.

```
<!-- composition_hasSubSequence_represents -->
<rule name="composition_hasSubSequence_represents">
    <premise>
        <subject   var="xxx"/>
        <predicate uri="&av;hasSubSequence"/>
        <object    var="yyy"/>
    </premise>
    <premise>
        <subject   var="yyy"/>
        <predicate var="propriete-representation-inconnue"/>
        <object    var="zzz"/>
    </premise>
    <premise>
        <subject   var="propriete-representation-inconnue"/>
        <predicate uri="&rdfs;subPropertyOf"/>
        <object    uri="&av;represents"/>
    </premise>
    <consequent>
        <subject   var="xxx"/>
        <predicate var="propriete-representation-inconnue"/>
        <object    var="zzz"/>
    </consequent>
</rule>
```

Code 5.6 – Déclaration de la composition de la propriété `hasSubsequence` et d'une propriété spécialisant `represents` dans **Sesame**

Conception d'une ontologie test

Pour tester les possibilités de chacun des éditeurs engagés dans l'évaluation, les organisateurs ont proposé de reprendre un thème introduit pour l'édition précédente de l'atelier, le voyage touristique. Pour cette évaluation, destinée à comparer les possibilités de modélisation des différents outils, les organisateurs n'avaient pas proposé d'ontologie toute faite – une telle ontologie aurait nécessairement reflété un point de vue de représentation particulier, introduisant certains types d'informations, et pas d'autres – mais un court texte présentant le domaine d'une application. Les ontologistes avaient ainsi une description naturelle des besoins liés à l'application (cf. figure 5.9), qu'ils pouvaient utiliser pour guider leur conceptualisation.

> ...
> We know that when a client makes a trip, he chooses : transport and accommodation. Hence, we start by determining the means of transport that are currently available for a travel agency. We will have in our ontology the following ones : planes, trains, cars, ferries, motorbikes and ships. There are no other kinds of transport.
> ...
> In fact, customers are usually interested in the kind of planes that they will fly on : Is it a Boeing, or is it an Airbus ? Furthermore, they are even interested in the specific model of the plane in which they will fly (a Boeing 717 or a Boeing 777).
> ...
> Concerning hotels, the agency recommends in all the cities : hotels, and Bed and Breakfasts. Hotels rank from 1 star hotels to 5 star hotels and each hotel belongs to one of these five categories. For all of them, the agency knows their facilities : address, telephone number, URL, capacity, number of rooms, available rooms, descriptions, dogs allowed, distance to the beach, distance to skiing, etc. The agency also knows the facilities of the rooms : number of beds, rates, TV available, Internet connection, etc.
> ...

Fig. 5.9 Extraits du texte donné pour la spécification des ontologies EON

En l'occurrence, le texte présentant les besoins de l'application est en fait assez court, mais peut donner lieu à un travail de conceptualisation suffisamment riche pour tester les possibilités des éditeurs en matière de représentation ontologique. Des entités comme `Reservation` ou `Means Of Transport` admettent de nombreuses spécialisations, ou impliquent l'existence de multiples liens entre les individus du domaine modélisé. Avec Véronique Malaisé, nous avons pu constater comment la structuration induite par l'utilisation des principes différentiels et l'introduction de concepts de haut niveau permettait de rationaliser et clarifier le résultat de l'acquisition des connaissances conceptuelles. Des notions extraites directement du texte (représentées en figure 5.10a) on parvient à une arborescence normalisée très organisée – la branche la plus profonde comporte 10 niveaux de spécialisation. Et ce, même si nous avons pour cette ontologie diminué la quantité et la complexité des notions de haut niveau. Ainsi, notre première distinction sépare les entités abstraites – catégorie dans laquelle nous avons détaillé les *types de données* au sens informatique – des entités concrètes. Celles-ci comprennent des entités temporelles – les différents actes de *réservation* – et des entités caractérisées par un rapport essentiel à l'espace. Cette catégorie regroupera à son tour des *objets géographiques*, des *artefacts* (les différents *moyens de transports*) et des objets biologiques (les diverses *personnes*), etc.

En revanche, la hiérarchie de relations ne comporte que des liens que l'on pourrait qualifier

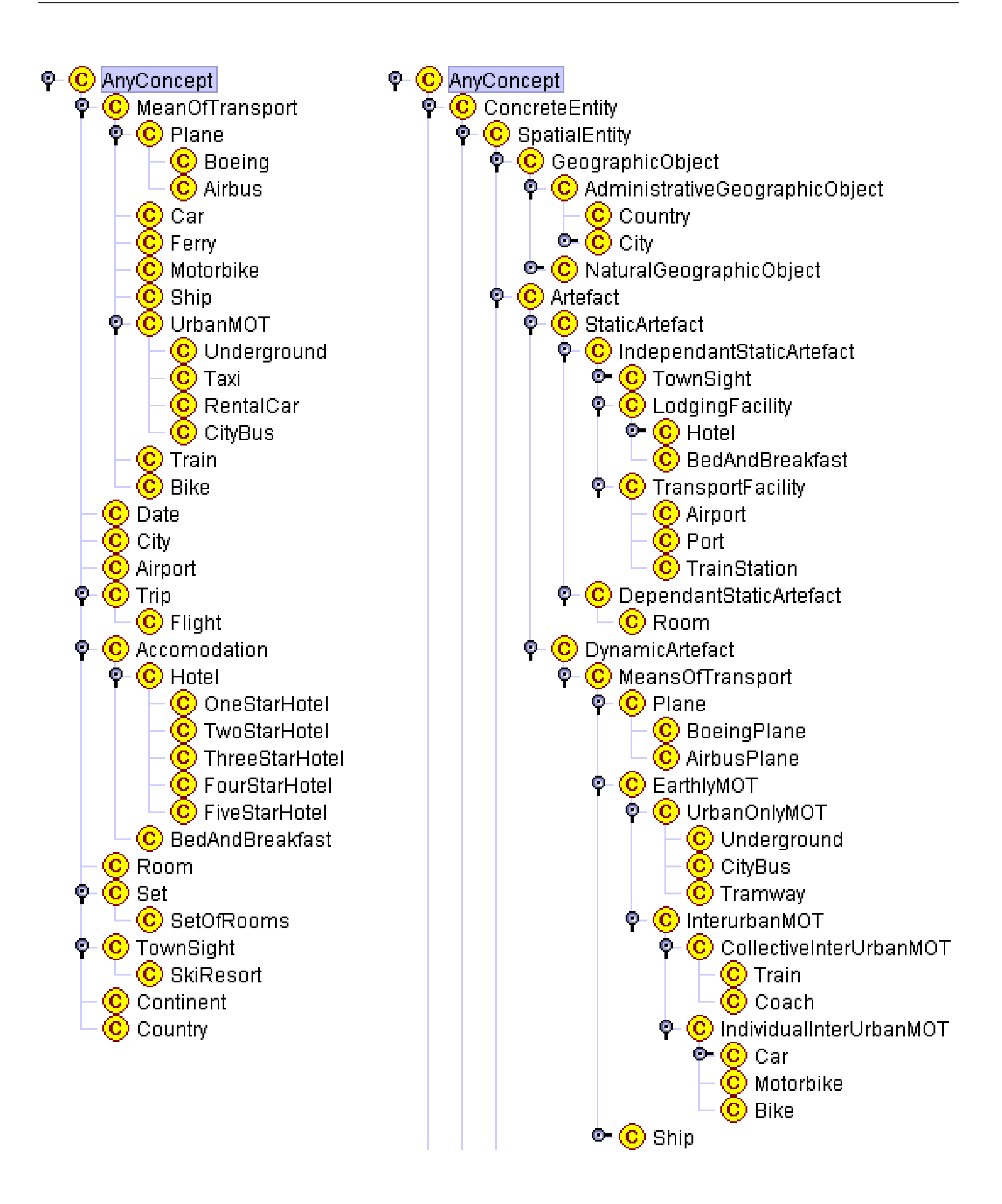


Fig. 5.10 Concepts de l'ontologie EON

(a) non structurés, (b) structurés par introduction de concepts de haut niveau et application de la méthodologie Archonte

d'*attributs* – pour des concepts comme `Room`, `TransportReservation`, etc. De tels attributs, attachés de manière privilégiée au concept dont ils permettent de caractériser les instances, sont peu propices à se spécialiser les uns les autres.

De fait, si l'ontologie comporte finalement 79 et 48 relations, il ne faut pas oublier qu'il s'agit d'une ontologie « jouet » dont l'objectif n'était que l'évaluation de notre outil **DOE**. Ainsi, si nous avons défini avec soin les domaines et co-domaines des relations[31], notre ontologie du voyage ne comporte pas de connaissances de raisonnement formelles, comme c'était le cas de certaines des ontologies réalisées pour EON 2002. Ultime conséquence du fait que cette ontologie ne soit qu'un exercice de modélisation, certes riche en enseignements (voir le contenu de la section suivante), elle ne présente pas de patron explicite qui mobiliserait ses entités en vue de la création d'assertions privilégiées dans un cadre applicatif précis.

Interopérabilité de DOE avec les autres éditeurs d'ontologies

Depuis le début de nos travaux relatifs à la mise en œuvre des principes méthodologiques repris dans cette thèse, nous n'avons pas placé notre outil de conception d'ontologies en concurrence avec les principaux environnements de développement proposés dans la communauté de l'ingénierie des connaissances. Il se concentre en effet sur la normalisation sémantique, alors que les autres visent tout particulièrement la formalisation et l'opérationnalisation, même s'ils s'efforcent par le biais d'interfaces graphiques élaborées, cachant la complexité des méta-langages sous-jacents, de rendre ces opérations plus intuitives pour le concepteur.

Un tel positionnement pour **DOE** nécessite cependant une articulation soigneuse avec ces autres environnements. Il faut pouvoir le cas échéant compléter le travail fait dans notre éditeur dans les outils qui permettent une formalisation et une opérationnalisation complètes, selon les critères définis par les langages standards de représentation d'ontologies. Il serait également souhaitable de pouvoir récupérer des ontologies déjà définies, mais qui ne suivent pas les principes méthodologiques avancés par Bruno Bachimont, pour pouvoir les re-travailler selon cette approche.

Pour cela, nous avons utilisé la technologie de transformation XSLT [XSL99] : le format de sauvegarde de notre éditeur utilise la syntaxe XML, il est donc possible de procéder relativement à des transformations – syntaxiques – de ce format en un autre format, en particulier si ce dernier utilise lui aussi cette syntaxe, ce qui est le cas des principaux langage d'encodage d'ontologies. Cette solution, si elle a contre elle le niveau auquel elle se place – on n'essaie pas vraiment d'interpréter la sémantique des objets manipulés au moment de la transformation – et sa piètre efficacité en termes d'utilisation des ressources machine, a l'avantage de l'adaptabilité. En effet, si la spécification d'un langage est modifiée, ou si un nouveau langage intéressant apparaît, ce qui peut arriver dans un domaine de recherche ou rien n'est fixé de façon définitive, il suffit de modifier la feuille de style correspondante, ou bien en ajouter une autre, sans avoir à modifier le code de l'outil de manière significative.

Lors de cette expérimentation, nous nous sommes intéressé à la façon dont DOE pouvait échanger des données avec les environnements **Protégé2000** (version 1.8), **OilEd** (version 3.5) et **OntoEdit** (version 2.6 libre)[32], comptant parmi les plus importants du domaine. En ce qui

[31] Il est à noter que l'expressivité de **DOE** nous a permis de représenter une relation ternaire, `isDistantFrom`, qui était utile pour exprimer de manière quantitative la distance entre deux entités spatiales, comme un aéroport et un hôtel.

[32] Cette expérimentation ayant eu lieu au début de notre thèse, ces outils ont considérablement évolué entre temps. Nous rappelons qu'une brève présentation de cet outil est présente dans ce manuscrit, dans la section 4.3.2 (page 128).

concerne le moyen d'échanger des ressources ontologiques avec ces environnements, notre choix a été d'étudier les possibilités offertes par les exports et imports RDFS puis OWL. Ces standards étaient effectivement pris en compte par la majorité des environnements de conception retenus, et par **DOE** lui-même.

RDFS comme langage d'échange En ce qui concerne l'export des informations de **DOE** vers les autres environnements, nos essais ont été plutôt concluants. Si RDFS ne permet évidemment pas d'encoder de manière explicite les principes différentiels, il est possible de les exporter automatiquement dans le champ `rdfs:comment`. Ce champ, destiné à contenir des informations textuelles, ne peut pas être structuré de manière à rendre compte explicitement des définitions différentielles. En configurant l'export de **DOE** de manière appropriée, on parvient tout de même à obtenir des commentaires lisibles, et surtout récupérables par tous les éditeurs de notre étude. Les informations de l'ontologie formelle sont également exportées avec succès : comme on l'a déjà fait remarquer, il y a des similitudes entre l'expressivité de RDFS et celle des supports GC pour l'export desquels notre éditeur a été conçu en premier lieu. On a donc là un résultat extrêmement important : on peut poursuivre le travail de formalisation et d'opérationnalisation entamé dans **DOE** dans un autre outil, en ayant toujours accès aux informations qui ont été définies antérieurement.

Les choses sont plus délicates en ce qui concerne l'import dans **DOE** d'ontologies créées avec d'autres outils. RDFS est en effet un langage qui, même dans sa syntaxe XML, admet plusieurs variantes syntaxiques, ce qui complique considérablement le travail de conception des feuilles de style XSLT. Par exemple, comme on l'a évoqué à la page 173, on peut déclarer une classe comme un élément `rdfs:Class` ou comme un élément `rdf:Description` de type `rdfs:Class`. De plus, comme ces ontologies sont essentiellement des ontologies munies d'une sémantique formelle, il peut être délicat d'essayer d'en extraire une structure différentielle. Le problème le plus important est celui de l'héritage multiple : d'un point de vue formel, aucun des parents d'un concept n'est privilégié, on ne raisonne qu'en termes d'inclusion d'ensembles. Dans ce cas, il faut donc choisir au hasard un des parents pour en faire le père de la notion différentielle correspondant à l'entité référentielle représentée en RDFS, ce qui est évidemment une solution non satisfaisante. On pourra se consoler en faisant observer que la phase d'export est beaucoup plus importante pour le positionnement de notre outil en *amont* des autres éditeurs, mais il demeure qu'on se retrouve là face à une limitation importante de notre outil, et, au-delà, de la méthode entière dont il s'efforce d'être la réalisation la plus fidèle. Pour que les significations différentielles soient complètement partageables par plusieurs éditeurs, il n'y a pas d'autre solution que de les intégrer dans le langage pivot, ou d'attendre qu'un langage de représentation soit suffisamment adaptable pour pouvoir être configuré en fonction des informations à représenter.

OWL comme langage d'échange Nos expérimentations avec OWL n'ont pas non plus été extrêmement concluantes, mais cela est dû en forte partie au fait que la moitié des éditeurs vers lesquels nous nous étions tournés ne le géraient pas encore. En effet, seul **Protégé2000** importait OWL. Puisque celui-ci avait déjà démontré sa capacité à importer nos ontologies traduites en RDFS, et que sur le plan de la syntaxe il n'y a que peu de différences entre la manière d'exprimer nos informations formelles limitées en RDFS ou en OWL[33], nous n'avons pas jugé utile de nous relancer dans de fastidieux tests. Cet unique résultat positif[34] n'a en effet été obtenu qu'après

[33] A part le fait, fondamental d'un point de vue sémantique mais minime d'un point de vue syntaxique, qu'une classe soit introduite en OWL par la primitive `owl:Class` et en RDFS par `rdfs:Class`...

[34] Ce qui ne diminue pas la valeur de ce résultat. Nous avons réussi à exporter nos ontologies à peu près correctement vers le seul éditeur capable à l'époque de les importer !

l'application de solutions quelque peu empiriques, que le lecteur intéressé pourra voir plus en détail dans notre article [ITM03]. Comme les expériences conduites pour essayer d'importer dans **DOE** des ontologies OWL produites avec les autres outils l'ont confirmé, l'absence de maturité – en tout cas à ce moment – de la norme avait pour conséquence des interprétations assez libres, nuisibles à l'interopérabilité des outils qui reposaient sur ces interprétations. Nul doute qu'à présent la même expérience aurait des résultats largement plus positifs.

5.2.5 Récapitulatif

Au cours de ces trois années de thèse, nous avons donc procédé à des expérimentations qui ont beaucoup apporté à nos réflexions, que ce soit sur le plan de la conception des ontologies ou de leur utilisation effective dans des systèmes d'indexation et de recherche. Cela nous a permis de tester nos hypothèses méthodologiques et nos outils, ainsi que les propositions des autres acteurs de la communauté de l'ingénierie des connaissances. Et même si la conception de certaines de ces ontologies ne s'inscrit pas directement dans le cadre de cas d'utilisation appliqués, il demeure qu'elles constituent une expérience concrète de modélisation. Ce qui en soi se révèle positif, d'autant plus que pour chacune des expérimentations – exceptée celle du voyage – nous nous sommes toujours efforcés de garder un lien relativement clair avec les usages observables dans notre domaine, celui de l'indexation.

La première des expérimentations, celle du cyclisme, s'inscrit en effet dans l'optique d'une assistance automatique à la création d'index. Pour cette expérimentation, qui ne demandait pas de création de système de description et de recherche en tant que tel, nous nous sommes surtout concentré sur la qualité de la conception de l'ontologie, en vue d'une exploitation de ses notions dans des mécanismes d'extraction d'information à partir de texte conçus par Estelle Le Roux. La proximité de ces notions avec les éléments de conceptualisation présents dans ces textes était impérative à la réalisation d'un système performant. En ce sens, la conception de cette ontologie a permis une validation qualitative de l'approche de conception par normalisation sémantique puis formalisation.

Ensuite, le cadre d'OPALES a permis une expérimentation complète de nos hypothèses, puisque le système d'indexation et de recherche conçu pour le projet avait pour ambition de permettre la création et l'utilisation de descriptions complexes dans des scénarios réels de recherche d'information. La proximité avec les usages, qu'il s'agisse de la reconnaissance d'éléments de conceptualisation pertinents ou de la facilitation de l'emploi des descripteurs dans des index exploitables utilement, était au cœur de ce projet. L'articulation de la normalisation sémantique avec les résultats des outils d'acquisition automatique de terminologie développés par Véronique Malaisé, la création de patrons d'indexation, la formalisation des ontologies faisant intervenir des connaissances de raisonnement réellement utiles, et finalement le contact avec des utilisateurs réels ont permis de valider qualitativement une grande partie de nos propositions. La taille des ressources ontologiques mises en jeu fait également de cette expérimentation une validation quantitative appréciable.

L'intérêt des expériences réalisées dans le cadre de l'atelier EON est moins immédiat. Cependant, si l'ontologie n'est pas de taille impressionnante et n'a pas été réellement utilisée, elle a constitué un terrain d'exercice dont l'apport est non négligeable pour le développement de notre savoir-faire en matière de conceptualisation. Qui plus est, elle a autorisé, *via* la comparaison du résultat d'une conception guidée par les propositions de Bruno Bachimont avec celui d'autres méthodes de conception, une certaine forme de validation qualitative. Elle aura également permis de préciser le rapport de notre outil de construction d'ontologies avec les autres solutions présentes dans la communauté, puis de tester la viabilité de notre positionnement. Et le résultat de

ce test fut positif, puisque finalement nous avons prouvé la faisabilité du seul mode d'interaction qui nous intéressait vraiment, à savoir l'export de nos ontologies vers d'autres environnements de conception.

Finalement, l'expérimentation réalisée dans le domaine médical en collaboration avec Raphaël Troncy aura été utile selon bien des angles. Tout d'abord, il faut noter qu'elle s'efforce de reprendre à son compte les préoccupations qui ont été dégagées d'OPALES et de l'étude des usages documentaires de l'INA, ce qui en fait un cas d'utilisation réaliste. Ensuite, elle a permis de tester la validité de nos hypothèses en les appliquant à l'aide des outils proposés dans le cadre de l'initiative du web sémantique, ce qui n'est pas négligeable vu que ce paradigme, même si ce n'est pas exactement le nôtre, est devenu la référence dans notre communauté. Enfin, la manière dont des ressources ontologiques d'une importance significative, qu'il s'agisse de réutilisation, d'adaptation ou de conception *ex nihilo*, peuvent somme toute être convenablement exploitées dans ce contexte technologique spécifique constitue une preuve de la valeur des hypothèses retenues, toutes ces ressources ayant à un moment ou un autre bénéficié de l'application de principes repris ou énoncés dans cette thèse.

5.3 Discussions méthodologiques

5.3.1 Utilisation d'ontologies pour indexer des vidéos

En ce qui concerne l'utilisation des ontologies, nos propositions – que l'on peut très brièvement regrouper en trois thèmes : indexer et rechercher dans des bases de connaissances, faciliter l'accessibilité des index, aider leur création par des graphes patrons – sont globalement validées par nos expérimentations. Trois problèmes importants subsistent cependant, sur lesquels l'orientation de nos réflexions n'a pas permis d'émettre des recommandations construites.

Articulation entre les fonctionnalités d'un SBC et celles d'un système documentaire

Au début de notre travail, nous avons fait l'hypothèse simplificatrice d'ignorer la différence entre ce que nous avons désigné comme des systèmes documentaires, à savoir des systèmes dont la fonction est la recherche des documents eux-mêmes, et les systèmes de recherche d'information, qui doivent renvoyer à l'aide des connaissances contenues dans une base des réponses factuelles aux requêtes des utilisateurs. Du point de vue des spécifications d'un SBC, et en particulier des connaissances ontologiques à définir et des inférences à conduire, cela ne semble pas faire de différence au premier abord. Et pourtant, la situation se complexifie très vite dans le cadre de systèmes comme OPALES qui ont à se préoccuper :

- du traitement des données assertionnelles – les index – par des mécanismes de raisonnement ;
- de la gestion des liens entre les index et les éléments documentaires eux-mêmes, ainsi que
- de l'exploitation des résultats que les moteurs d'inférence permettent d'obtenir à partir de ces index.

Par exemple, à partir des deux indexations – à vocation purement pédagogique – présentées en figure 5.11, on peut considérer trois cas de figure.

recherche documentaire classique. Dans ce paradigme, l'accent est mis sur la recherche de segments documentaires – des documents dans leur intégralité ou bien leurs parties – dont le contenu est pertinent par rapport au besoin informationnel de l'utilisateur, besoin dont la requête est l'expression. Ici, chaque index est considéré comme une base de faits à lui tout seul ; on ne peut

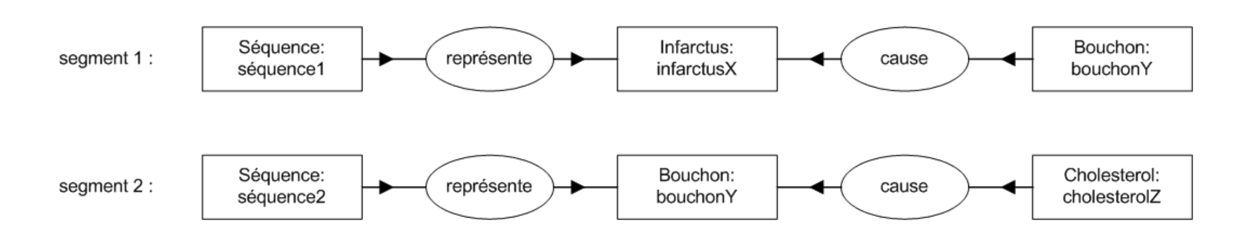


FIG. 5.11 Deux index fabuleux

se servir des connaissances apportées par les autres index pour répondre à une requête. Ainsi, si le segment 1 pourra répondre à une recherche de séquences représentant un accident causé[35] par un problème artériel, puisque les concepts d'infarctus et de bouchon spécialisent ces deux notions. En revanche, il ne pourra pas figurer dans les résultats d'une recherche de séquences représentant des accidents causés par des anomalies sanguines, puisque la connaissance selon laquelle le bouchon est lié à un taux élevé de cholestérol est présente dans un autre index.

recherche d'information. Dans ce paradigme, le lien entre la connaissance représentée par l'index et le document lui-même est mis au second plan : on recherche un contenu informationnel avant tout. Ici, tous les index sont considérés comme une seule base de faits. On saura donc que `cholesterolZ` a causé `infarctusX`, ce qui permet de répondre à des requêtes cherchant toutes les causes d'accidents cardiaques apparaissant dans la base d'index.

Pour obtenir des résultats qui soient vraiment des documents, il faut ajouter des fonctionnalités strictement documentaires, extérieures au système de raisonnement lui-même. Le segment 1 pourra selon toute logique répondre à une recherche de séquences représentant un accident causé par une anomalie sanguine. Il est cependant important de mentionner que pour cela *il faut que le système le lie correctement à sa représentation dans son index même.* Ceci ne se passera que si on se place dans un cadre tel que celui proposé dans la thèse de Raphaël Troncy [Tro04], que nous avons repris dans notre expérimentation commune présentée en section 5.2.3. En l'occurrence, il faut créer une représentation des documents à l'aide d'un langage documentaire comme MPEG-7 [MPEG-701], qui permettra d'introduire par exemple un découpage du document indexé en séquences. Le système de représentation des connaissances devra ensuite contenir des instances de concepts tels que le concept `Sequence` de l'ontologie de l'audiovisuel – qui seront explicitement reliés aux représentations purement documentaires que l'on vient d'évoquer. Il faut noter que bien que le point de vue « eau » d'OPALES utilise de tels concepts documentaires, il ne permettrait pas un tel fonctionnement, puisque les index ne sont pas mis en commun, et qu'il n'existe pas dans le système de lien explicite entre un segment d'un document et sa représentation dans le SBC.

recherche documentaire évoluée. On peut cependant vouloir des systèmes de recherche documentaires qui vont plus loin dans l'articulation entre documents, index et raisonnements. De fait, si l'on effectue une recherche sur les liens entre un accident cardiaque et une anomalie sanguine, nos deux segments pourraient être renvoyés. Mais il serait encore plus intéressant de les renvoyer en une seule *combinaison* documentaire – ne serait-ce qu'une simple concaténation des deux segments – qui soit considérée comme une unique réponse au besoin exprimé. Le problème est qu'une telle solution demanderait dans une certaine mesure l'application de stratégies

[35] Pour les besoins de notre argumentation, nous considérons que la relation `cause` est transitive.

d'explication des raisonnements conduits. Il faut en particulier conserver les liens avec les index ayant permis la conduite des raisonnements. Il serait également nécessaire de faire appel à des connaissances indiquant le moyen dont un système peut composer de manière semi-automatique de nouveaux documents à partir des raisonnements conduits et des liens entre les résultats de ces raisonnement et les sources documentaires qui l'ont permis. De telles « méta-connaissances » de composition devraient sûrement être mises en rapport avec les connaissances ontologiques du point de vue thématique retenu pour l'indexation : dans notre exemple, il serait intéressant d'introduire des contraintes relatives au lien causal, pour déduire que dans la concaténation le segment 2 doit être placé avant le segment 1...

Evidemment, le dernier mode semble le plus intéressant du point de vue de la recherche sur les systèmes documentaires, puisqu'on approcherait alors de très près le travail d'un humain[36]. Il est cependant probable que l'importance du travail de modélisation requis, ainsi que les capacités de raisonnement actuelles mettent hors d'attente la réalisation de projets semblables à l'heure actuelle. Et quelques travaux d'automatisation des processus d'édition documentaire à l'aide des connaissances sur les contenus apportent des pistes intéressantes, ils utilisent des schémas de publication non dynamiques, dont l'instanciation ne dépend aucunement des raisonnements qui ont été conduits pour trouver les éléments qui s'insèreront dans ces schémas [Ran00, RBv$^+$00, GBvH03].

Quels raisonnements pour les systèmes d'indexation à base de connaissances ?

Tout au long de notre travail, nous avons insisté sur l'importance de l'inférence comme moyen d'assister les pratiques documentaires de ceux qui utilisent le système. Cependant, en la matière, de nombreux choix restent possibles, qui sont fonction de la richesse des connaissances que l'on souhaite spécifier, mais aussi des besoins en ce qui concerne les performances des SBC. Comme on peut s'y attendre dans une section de discussion, nous n'allons pas proposer de solution générale au problème du choix, même si, en ce qui concerne l'indexation, nous avons pu voir que certaines pistes sont à privilégier...

De fait, il existe des études qui essaient d'analyser l'expressivité des langages de représentation de connaissances et des règles d'inférence qui doivent permettre une manipulation optimale de ces connaissances. Dans le cadre du web sémantique, qui essaie de standardiser des langages et des outils adaptables à un maximum d'applications tout en conservant des propriétés permettant leur emploi dans les conditions d'échelle et de performance requises pour le web, ces études prennent un intérêt tout particulier. De fait, celles-ci démontrent souvent que les possibilités offertes par ces standards sont insuffisantes, et poussent à l'adoption de standard plus évolués. Ainsi, [GDGB03] présente une analyse des besoins en matière de conception et d'utilisation d'ontologies biomédicales. Et en déduit que si les fonctionnalités offertes par les langages de spécification du web sémantique et les outils[37] qui les mettent en œuvre sont utiles, il peut être nécessaire de se tourner vers d'autres langages pour représenter des connaissances de raisonnement pertinentes pour les domaines applicatifs concernés, notamment pour définir des règles de composition de relations.

Des solutions adaptées, mélangeant techniques standard du web sémantiques et outils issus du champ « traditionnel » de l'intelligence artificielle, peuvent être développées pour répondre

[36] Ce pourquoi les travaux existants visent dans un premier temps l'élaboration d'assistant à la publication, plutôt qu'une création documentaire entièrement automatisée.

[37] En l'occurrence, OWL, **Protégé** et des raisonneurs pour les logiques de description, tels que **FaCT** [Hor98] et **Racer** [HM03].

de façon ponctuelle à de tels problèmes [GI04], en attendant éventuellement l'émergence de nouveaux standards. Mais ces approches peuvent parfois souffrir du manque de maturité des systèmes employés, ou tout simplement de la trop grande complexité des algorithmes requis. Dans un contexte proche du nôtre – celui de la caractérisation automatique de régions de documents visuels – [LH04] montre bien les problèmes de performances rencontrés en utilisant des outils qui font encore partie de l'état de l'art en matière de recherche. On retrouve donc le dilemme traditionnel de la représentation formelle de connaissances et de leur manipulation automatique par des mécanismes de raisonnement. Recourir à des connaissances et des traitements riches est évidemment intéressant au regard de la qualité des résultats obtenus, mais les efforts requis sont plus importants, que ce soit pour les concepteurs des ressources ontologiques qui seront utilisées par le systèmes, pour ceux qui créent les assertions factuelles qui représentent les contenus documentaires, mais aussi pour le système qui effectue les traitements automatiques.

Or, il semble qu'un haut degré de complexité ne puisse faire l'objet d'économies pour les applications qui nous intéressent. En ce qui concerne les besoins spécifiques à la description de contenus documentaires, et en particulier de contenus audiovisuels, nous ne sommes en effet pas les seuls à avoir dégagé la nécessité des connaissances de raisonnement illustrées au fil de ce manuscrit. Même s'ils ne s'inscrivent pas dans une approche méthodologique comparable à la nôtre, et s'attachent moins au cas du document audiovisuel, des travaux comme ceux exposés dans [GHB96] et [Oun98] ont dégagé l'importance des connaissances relatives aux relations, et en particulier à l'agencement de celles-ci.

Une telle approche semble tout à fait envisageable dans le cas de volumes d'information relativement réduits : après tout, le comportement du moteur d'inférences d'Opales n'a pas montré de signe de faiblesse, pas plus que les systèmes que nous avons employés dans notre expérimentation de la section 5.2.3. Cependant, des études quantitatives impliquant un plus grand nombre de données seraient intéressantes, si l'on veut par exemple élargir le champ de nos propositions à des collections documentaires plus importantes.

Ce problème de la complexité des mécanismes d'inférence peut être davantage obscurci par le fait que certains travaux proposent de se tourner vers des systèmes qui sortent du cadre bien étudié de la logique du premier ordre. [MCMO03] propose par exemple de revenir vers des techniques de recherche d'information traditionnelle, où le calcul de l'implication logique est remplacé par des considérations sur le poids des concepts et relations dans un index particulier, et dans une collection d'index. Une telle technique présente l'avantage d'autoriser un classement des résultats d'un requête par ordre de pertinence[38]. D'autres interprétations de l'« implication » partagent cette préoccupation : l'utilisation de mécanismes de calcul de *degré de similarité* entre représentations, en s'aidant des informations fournies dans les ontologies – ou en tout cas dans les hiérarchies conceptuelles que l'on y trouve – est une piste que l'on rencontre parfois [MGLB01, AHAS03, ANN03]. Plus proche de la déduction logique traditionnelle figurent les solutions recourant à la *logique floue*, ou les éléments d'un index sont associés à des niveaux de certitude [TBH03] qui seront ensuite exploités par des mécanismes de déduction appropriés. Ces approches sont cependant à l'heure actuelle adoptées par un très petit nombre de chercheurs, ce qui ne leur garantit pas une pérennité comparable à celle des solutions plus classiques, qui bénéficient des efforts de communautés plus importantes.

Mais la multiplicité des choix possibles en matière de raisonnement pour la recherche d'information peut ne pas remettre en cause les choix fondamentaux que nous avons faits. S'il existe évidemment des travaux qui ignorent la dimension relationnelle lors des raisonnements, ce n'est

[38] Alors que l'implication logique permet seulement de décider qu'un document est pertinent ou ne l'est pas.

pas le cas de ceux que nous venons de citer. Et l'on peut envisager l'adaptation de nos propositions à de tels cadres non standard, même si celle-ci n'est pas évidente. Par exemple, la notion de patron d'indexation peut être maintenue, quand bien même ledit patron serait « flou ». Et si le calcul de similarité dans un contexte non logique est une proposition essentiellement différente du recours à des axiomes spécifiant des cadres de déduction logique, on peut faire observer que la première approche peut très bien répondre elle aussi au problème fondamental de compensation de la variabilité des index soulevé par l'exigence de continuité sémantique.

Faut-il utiliser une ontologie par défaut pour l'indexation audiovisuelle ?

En fait, pour dégager les types de raisonnements requis pour les systèmes d'indexation, certains des travaux d'analyse que nous venons d'évoquer se basent sur une certaine classe de connaissances, à savoir celles qui sont liées directement aux documents considérés et à leurs propriétés les plus évidentes, ainsi qu'aux implications de la tâche particulière que constitue l'annotation ou d'indexation de ces documents. En effet, comme on l'a vu dans le chapitre 1, la description d'un document suppose toujours la *localisation* d'un segment documentaire particulier, la *qualification* de ce segment à l'aide du langage d'indexation retenu, et la *structuration* des unités documentaires ainsi décrites en un ensemble rendant compte de leur agencement. Quel que soit le contenu de la qualification – à orientation plus *thématique* –, les deux autres phases nécessitent l'appareillage de systèmes *documentaires* complexes, qui peuvent faire appel à des connaissances et des vocabulaires dédiés. Ce fait, déjà reconnu depuis longtemps dans les approches documentaires classiques, en particulier pour le cas audiovisuel [Pic96, Mou99], peut justifier en soi le recours à des ontologies et des procédures de raisonnement spécialisées dans une approche plus formelle [DB00].

Quelques-unes des approches logiques de la recherche d'informations que nous avons déjà présentées[39] proposent donc des vocabulaires de descriptions minimaux ainsi que des *contenus* axiomatiques par défaut qui correspondent à des fonctionnalités strictement documentaires. Des travaux comme ceux du laboratoire MRIM [MBC95, CMF96] avancent ainsi des primitives formelles permettant de construire un langage logique d'indexation de l'image. Ces primitives sont accompagnées d'un certain nombre d'axiomes qui permettent de spécifier le comportement du système de recherche d'information documentaire qui les emploiera. Dans un contexte plus explicitement ontologique et « audiovisuel », Raphaël Troncy a montré qu'il était souhaitable d'articuler l'ontologie du domaine traité par les documents que l'on veut indexer à une ontologie permettant de représenter ces documents eux-mêmes [Tro04]. Certaines tâches habituellement dévolues aux systèmes documentaires peuvent en effet être prises en charge de manière plus appropriée par un système à base de connaissances[40].

Cette problématique de la constitution d'un ensemble de connaissances documentaires par défaut n'était pas au cœur de notre thèse. Comme dans le cas d'OPALES, nous aurions pu nous contenter d'une indexation purement thématique, et laisser à un système documentaire le soin de gérer le lien entre ces index et les documents qu'ils décrivent. Dans une telle vision, un « point de vue ontologique documentaire » ne serait pas essentiellement plus important que les autres. Cependant, force est de reconnaître que ce point de vue est extrêmement important, surtout dans le contexte de l'INA. Dans la lignée des travaux de recherche que notre département effectue il est en effet tout à fait raisonnable de penser que c'est bien le rôle de notre Institut que de produire une ontologie qui convienne à son « métier ». De fait, un tel sujet correspond à

[39]Notamment en section 2.4, pages 66 et suivantes.

[40]Il faut mentionner que l'on retrouve ici une partie des problèmes que l'on a déjà évoquées dans cette section, page 187.

des pratiques relativement bien établies, qui créent un cadre satisfaisant d'activité d'acquisition des connaissances. Et finalement, d'un point de vue plus personnel, il s'agissait là d'un terrain d'expérimentation extrêmement intéressant pour nos propres hypothèses

Ainsi que nous l'avons mentionné dans la section 5.2.3, cette ontologie, à la constitution de laquelle nous avons très activement participé, a été inspirée par les usages de l'INA. Sous notre impulsion, et au contact de chercheurs spécialistes de la sémiotique du document audiovisuel, comme Peter Stockinger [Sto03a] ou Louis Chamming's de la direction de la recherche de l'INA, elle a pris un tour plus « sémiotique ». Elle contient a présent des primitives permettant, comme dans le cas du point de vue audiovisuel d'OPALES, de faire la distinction entre la description d'une forme audiovisuelle, celle de son contenu thématique[41] et celle de la *liaison* entre les deux niveaux précédents. *In fine*, on dispose d'un contenu ontologique propre aux spécificités du document audiovisuel, *auxiliaire* dans la mesure où il devra être associé à différentes ontologies de domaines pour les besoins des applications précises que l'on aura identifiées.

5.3.2 Faciliter la conception des ontologies

En matière de méthodologie de conception d'ontologies, beaucoup de points restent également encore en suspens, qui pourraient perfectionner nos propositions. Nos efforts de reprise et d'instrumentation de la méthodologie de conception de Bruno Bachimont n'ont en effet pas résolu tous les problèmes que pourraient se poser un lecteur exigeant. De même, l'articulation que nous proposons entre patrons de conception et patrons d'utilisation ontologiques peut soulever de nouvelles questions.

Implémentation de la méthodologie ARCHONTE

La méthodologie proposée par Bruno Bachimont constitue un cadre particulièrement riche. On a vu en particulier comment le recours à la normalisation sémantique pouvait guider tant la manière de concevoir les ontologies que celle de les utiliser. Néanmoins, lorsqu'on essaie de la mettre en pratique, de nombreuses questions apparaissent. Nous avons apporté des réponses à plusieurs d'entre elles dans nos travaux, mais certaines n'ont pas été traitées : par manque de temps, ou bien tout simplement parce qu'elles sortaient des objectifs de départ de notre travail de thèse.

l'articulation du niveau du domaine applicatif avec des notions de haut niveau. On a vu, au cours des différentes parties de ce manuscrit consacrées aux expérimentations de conception ontologique, qu'il faut s'efforcer de rattacher les primitives conçues pour une application donnée à des notions plus abstraites, bénéficiant d'un effort de légitimation théorique conséquent. Ce besoin, que nous avons affirmé clairement dans le cadre de la réutilisation de patrons de conception, a été quelque peu nuancé dès lors qu'il s'agissait de s'engager par rapport à une structure lourde, sans rapport immédiat avec l'usage visé. De fait, les contraintes imposées par la normalisation sémantique – et notamment l'interdiction de l'héritage multiple – forcent le concepteur de l'ontologie à un engagement extrêmement fort dans l'étape de normalisation. Or il peut adapter par la suite certains des choix qui ont été faits lors de cette étape, la formalisation des connaissances autorisant une flexibilité plus grande qui permet de se rapprocher des conceptualisations pragmatiques que l'on rencontre dans les domaines d'applications concrets.

[41] Ce contenu peut lui-même être représenté de manière complexe, comme nous l'avons recommandé, mais cela devra se faire avec l'aide d'une ontologie thématique appropriée, et non de celle de l'ontologie de l'audiovisuel elle-même.

Il faut donc trouver un moyen de rationaliser les concepts différentiels de plus haut niveau de manière à produire l'engagement sémantique le plus clair possible, de sorte à ce que le concepteur puisse le respecter tout en identifiant clairement les marges de manœuvre dont il peut disposer. Cela implique un travail d'étude et de conceptualisation important, qui peut se concrétiser en particulier par :
- une réutilisation plus critique des ressources existantes, en particulier de l'ontologie Menelas, dans l'optique d'évaluer leur adaptabilité à des applications concrètes et diverses ;
- une tentative de normalisation des différentes ontologies de haut niveau proposées à un niveau formalisé, comme les propositions de John Sowa [Sow00], celles du groupe de travail SUO[42], l'initiative GOL [DHHS01] ou bien évidemment l'ontologie DOLCE.

dualité des connaissances ontologiques et représentation dans un langage standard. Lors de l'expérimentation EON, nous avons vu qu'il était important de pouvoir transférer les informations spécifiées lors de la normalisation sémantique dans les formats de représentation les plus utilisés. Cela est indispensable si l'on veut continuer la formalisation et l'opérationnalisation au-delà des fonctionnalités élémentaires que **DOE** autorise. Et nous avons montré que cela est possible. Pour autant, en l'absence de standard à même de répondre à nos préoccupations, l'export de ces informations ne peut se faire que de façon détournée...

En fait, notre souci rejoint un problème plus vaste, qui n'a été abordé que très peu dans les efforts méthodologiques en rapport avec la conception des ontologies. Celles-ci sont en effet destinées à être représentées le plus souvent *via* des langages standard dont les capacités sont plus tournées vers la spécifications de calculs d'inférence que vers la représentation de connaissances relatives à la signification générale des notions de l'ontologie. Ces langages sont certes supposés mettre en œuvre des mécanismes de définition relativement naturels, comme le recours à la relation de subsomption. Mais nous avons vu qu'il était préférable d'ajouter des informations qui, par exemple parce qu'elles ne sont pas formalisées, ne seront pas directement utilisables par un SBC. Dans l'absolu, un processus d'acquisition des connaissances ontologiques aboutirait donc à la création de deux sortes de connaissances :
- des connaissances formelles et opérationnelles, directement en rapport avec l'exploitation de l'ontologie par les mécanismes de traitement automatiques d'un SBC ;
- des spécifications auxiliaires, établies lors de l'acquisition des connaissances pour guider la tâche des agents humains interagissant avec l'ontologie pendant la conception ou l'utilisation de celle-ci.

La deuxième classe de connaissances est plutôt variée. On y retrouve les principes différentiels de la méthodologie avancée par Bruno Bachimont, mais aussi les résultats de l'analyse ontologique *formelle* proposée par Nicola Guarino. De fait, les informations que ces connaissances contiennent peuvent être informelles comme formelles : les méta-propriétés de la méthodologie ONTOCLEAN sont effectivement formalisables à l'aide de la logique modale, logique qu'a également utilisée Bruno Bachimont dans [Bac01] pour proposer une formalisation des informations différentielles. Mais ces informations formelles peuvent être hors de propos pour les applications utilisant l'ontologie. Elle risquent même parfois de faire intervenir des entités qui ne pourront jamais être représentées à l'aide du vocabulaire de l'ontologie, faut de motivation applicative pour introduire les concepts et relations correspondants. Par exemple, dans l'ontologie du cyclismes, nos lieux géographiques administratifs sont *dépendants* d'une activité administrative, du point de vue de l'analyse ontologique *formelle*. Pourtant, une telle activité a fort peu de chances d'apparaître un jour dans une base de connaissances représentant des événements cyclistes...

[42] *Standard Upper Ontology Working Group*, `http://suo.ieee.org`.

Le problème est que ces connaissances de niveau « méta » ne vont donc que pouvoir être difficilement représentées dans les langages formels classiques de représentation d'ontologies. Dans un cadre d'acquisition des connaissances « traditionnel », une ressource conceptuelle reste employée dans un contexte limité, le plus souvent celui d'un système d'information spécifique ou d'une organisation particulière. Pour les ontologies, surtout si celles-ci sont créées en conformité avec la vision du web sémantique, la situation est différente. Leur représentation se doit en effet d'incorporer toutes les informations permettant de cerner la conceptualisation qu'elles représentent, surtout si ces informations ont fait l'objet d'une explicitation lors de la phase de conception. Cela est indispensable au partage des compréhensions associées aux primitives contenues dans ces vocabulaire, cela peut aussi l'être dans le cas où des systèmes auraient à accéder à des éléments de signification qui n'avaient pas d'utilité évidente pour les taches classiques de description et de recherche d'information. Valentina Tamma, qui dans sa thèse a proposé un jeu de méta-propriétés relationnelles [Tam02], insiste bien sur le fait que de telles informations pouvent très bien servir comme données pour des algorithmes d'alignement d'ontologies. Elles sont par conséquent susceptibles d'améliorer l'interopérabilité de SBC reposant sur des conceptualisations différentes, ce qui est l'un des objectifs de l'initiative du web sémantique.

Pour notre éditeur d'ontologies, qui devait évidemment enregistrer et accéder à l'ensemble des informations requises par la méthodologie ARCHONTE, nous avons créé un format de représentation qui encode explicitement les volets différentiel et formel des ontologies. Il faudrait voir comment on peut prolonger une telle approche, sans alourdir plus qu'il n'est nécessaire les standard existants. L'opportunité offerte par une note de travail concernant le langage OWL[43] de créer des *annotations structurées* répondant à des schémas prédéfinis semble en particulier être intéressante. Celle-ci permettrait en effet à ceux qui le désirent de créer des outils capables d'interpréter des informations qui demeurent optionnelles, le cœur des spécifications formelles restant à la portée de tous les outils qui se conforment au standard.

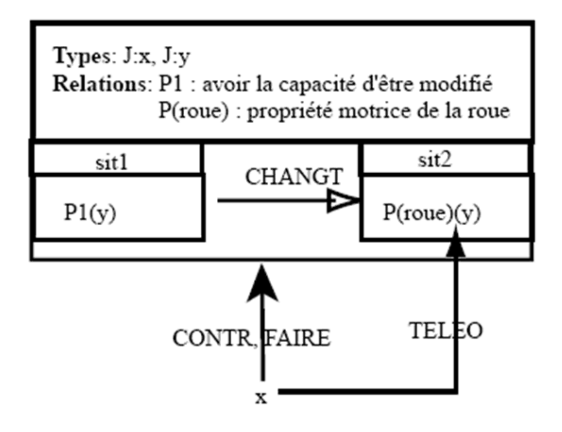


FIG. 5.12 Un schème sémantico-cognitif du verbe *rouler*, extrait de [Dji00]

définition des connaissances ontologiques de normalisation sémantique. Finalement, pour mentionner un aspect qui sortait clairement des objectifs initiaux de cette thèse, on pourrait s'attarder sur la nature même des informations qui permettent d'obtenir la normalisation

[43]Il s'agit d'une proposition de *syntaxe de présentation*, reposant sur XML, et différente de la syntaxe RDF qui s'est imposée depuis. Cf. `http://www.w3.org/TR/owl-xmlsyntax/`.

sémantique. La sémantique différentielle est-elle la seule à autoriser une normalisation ? Nous venons en effet d'un laboratoire universitaire – le Lalicc – qui propose une théorie sémantique qui s'intéresse à la détermination de types primitifs, puis à leur utilisation au sein de structures qui rendent compte des significations des éléments de la langue en une démarche cohérente et formalisée [Des90]. On trouve notamment des *schèmes sémantico-cognitifs* [Abr95] (voir figure 5.12) qui dans le domaine verbal permettent de définir par exemple des situations statiques, ou bien des évolutions qui s'appliquent aux différentes entités impliquées par les actions d'énonciation.

De tels travaux seraient d'un intérêt certain pour épauler l'analyse ontologique d'un domaine d'application. De fait, nous avons essayé de les prendre en considération lors de quelques-unes de nos expérimentations, comme nous l'avons évoqué en section 5.2.1. Cependant, force est de reconnaître que nos efforts ont été relativement peu satisfaisants d'un point de vue scientifique. Si les primitives sémantiques de cette théorie sont adaptées à la conception de vocabulaires abstraits cohérents, le passage au niveau d'un domaine d'application peut s'avérer dans un certain sens contre-productif. Les propositions qui sont faites ici prennent en effet tout leur intérêt dès lors qu'il s'agit d'exprimer les nuances des expressions en langue générale – la polysémie verbale, par exemple –, et non pour rendre compte de conceptualisations extrêmement marquées par leur contexte d'application, qui nécessitent un effort d'une autre nature, bien moins *réaliste*[44].

De plus, il n'est pas dit qu'une telle théorie sémantique réponde à nos impératifs en matière d'accessibilité. Comme on l'a vu, le travail d'adaptation des ressources conceptuelles complexes à des cas applicatifs simples peut se révéler délicat, et des utilisateurs non experts pourraient ne pas être capables d'en saisir les subtilités. De fait, nous ne fermons évidemment pas la porte à l'emploi de ces solutions pour un cadre ontologique appliqué. Mais une telle piste demande plus de réflexions que ce que nous avons pu lui consacrer pendant cette thèse CIFRE, qui devait se concentrer sur des problèmes plus immédiats.

Il en va de même pour d'autres théories sémantiques, comme celle de James Pustejovsky [Pus01]. Celui-ci propose en en effet une décomposition des significations lexicales en *qualia* (cf. figure 5.13) qui représentent les attributs fondamentaux des éléments définis. Mais là encore, comme l'article que nous venons de citer le présente d'ailleurs fort bien, ces travaux ciblent prioritairement la conception de ressources générales. Leur adaptation à des cas particulier – et ses conséquences sur le statut des définitions apportées – reste encore à étudier.

$$\begin{bmatrix} \textbf{beer} & \\ \text{ARGSTR} : & \begin{bmatrix} \text{ARG1} : \texttt{x:liquid} \end{bmatrix} \\ \text{QUALIA} : & \begin{bmatrix} \text{FORMAL} : \texttt{x} \\ \text{TELIC} : \texttt{drink(e}^{P}\texttt{,y,x)} \end{bmatrix} \end{bmatrix}$$

Fig. 5.13 La structure de *qualia* du nom commun « beer », extrait de [Pus01]

Il serait donc possible de trouver des sémantiques linguistiques alternatives à la sémantique interprétative de François Rastier, à la condition de pouvoir conserver la faculté qu'a cette théorie de pouvoir justement fournir des mécanismes qui s'adaptent à des contextes interprétatifs précis. Le fait que toutes ces définitions soient formalisées et complexes est une raison supplémentaire de tester leur réutilisabilité. A un niveau générique, nous tenons peut-être là des structures dont on pourrait faire dériver des patrons de conceptions ontologiques. . .

[44]Au sens où dans notre optique, plus *pragmatique*, le *biais* applicatif peut l'emporter sur la création d'un système cohérent permettant de rendre compte d'un nombre maximal des phénomènes observables.

Les patrons dans la conception des ontologies

des légitimations « naturelles » pour les patrons. Jusqu'à présent, les patrons d'indexation dont nous avons parlé ont été obtenus par acquisition manuelle des schémas de connaissances structurants du domaine applicatif considéré. Le problème est que si on réutilise des patrons de conception qui légitiment nos efforts d'un point de vue théorique, on n'a pas créé le lien direct avec les pratiques attestées – via et les textes ou gloses qui peuvent les encoder – que l'on avait obtenu en ce qui concerne la substance des descriptions, à savoir les concepts et les relations que les patrons mobilisent.

Il serait donc intéressant de justifier la forme de ces patrons, elle aussi, à l'aide de théories sémantiques adaptées. Toute forme d'assistance au travail manuel d'acquisition de cette forme serait également la bienvenue. Comme on vient de le faire remarquer, on pourrait réutiliser les résultats des investigations des équipes de Jean-Pierre Desclés ou James Pustejovsky, puisque ceux-ci proposent des *dictionnaires* associant des significations structurées aux éléments de la langue.

Plus prosaïquement, réussir à isoler puis rattacher nos patrons à des expressions langagières issues de corpus relevant du domaine de l'application ciblée serait déjà une avancée appréciable. Comme nous l'avons mentionné auparavant, des travaux existent, qui sont capables d'explorer les textes pour extraire des candidats pour les concepts et relations d'une ontologie [BAC04b]. La plupart s'attachent au repérage des définitions pour associer des réseaux à des éléments conceptuels considérés comme centraux [Le 00, Mal05]. Dans l'optique de l'extraction de patrons d'indexation, il faudrait pouvoir adapter ces outils pour qu'ils puissent repérer des schémas d'*utilisation* de ces entités dans le domaine applicatif. Au cours de cette thèse nous avons commencé à réfléchir à de tels objectifs en compagnie de Véronique Malaisé, doctorante à l'INA. Mais nous n'avons pu y consacrer les efforts qu'à nos yeux cette piste mérite.

articulation entre patrons de conception et patrons applicatifs, économies cognitives et engagement ontologique. Lorsqu'on veut rendre compte des conceptualisations existant dans un domaine particulier, on se retrouve souvent face à des idiomes propres à ce domaine, marqués par l'usage de figures de style ou bien de glissements de sens qui reflètent des façons de penser profondément ancrées dans les pratiques.

Ainsi, dans le domaine du cyclisme, on emploiera la préposition « avant » dans un sens qui n'est pas immédiatement temporel : « Il a chuté avant le col » est une expression attestée. Or, pour représenter une telle connaissance, il faudrait en première approche lier par une relation exprimant l'antériorité temporelle un *événement* – la chute – et un objet *spatial* – le col. Ceci serait évidement incorrect d'un point de vue ontologique. En revanche, on peut faire remarquer que cette pratique est autorisée dans le domaine à cause de la spécificité de la notion de course, qui se déroule sur un *parcours*. Un repérage temporel entre des événements peut en effet impliquer une forme de repérage « spatio-temporel » sur l'axe du trajet de la course, que le parcours temporel munit d'une orientation bien définie. On peut même créer un axiome[45] rendant compte d'une certaine forme d'implication entre ces deux relations de repérage :

$$\forall t,l\ (ObjetTemporel(t)\ \wedge ObjetSpatial(l))\ \rightarrow\ [\ avantST(t,l)\ \rightarrow$$
$$\exists u\ (ObjetTemporel(u)\ \wedge localisationSpatiale(u,l)\ \wedge avantT(t,u))\ ]$$

L'articulation des patrons d'utilisation avec les patrons de conception permet justement de mettre en lumière – et même de prendre en compte automatiquement – certaines de ces économies

[45] Les noms donnés ici aux concepts et relations ne reprennent pas nécessairement ceux de nos ontologies.

cognitives spécifiques à un domaine. En effet, les patrons de conception sont génériques, ils ne subiront donc pas eux-mêmes ce type d'adaptation[46]. Il peut donc être intéressant de situer un engagement ontologique par rapport à de telles références. Par exemple, dans le domaine de l'audiovisuel, le patron d'utilisation de notre ontologie (cf. figure 4.7, page 145) semble laisser de côté nombre des distinctions introduites par le patron *D&S*, distinctions que l'on avait pourtant reprises lors d'une première spécialisation de ce patron de conception (cf. figure 4.6, page 144). En particulier :

– les *rôles* ne sont pas séparés des *objets* qui les jouent, mais sont directement introduits pour classer ceux-ci,
– les *valeurs* des *paramètres* attachés aux rôles dans le patron de conception sont directement liés aux objets, *via* des attributs de production ou de diffusion.
– les activités et séquences d'événements ne sont plus présents dans le patron d'utilisation.

Les deux premiers points correspondent à des économies de représentation classiques, que combattent justement les chercheurs qui ont introduit des critères de modélisation comme ceux à l'œuvre dans DOLCE. Cependant, force est de reconnaître que cette façon de faire se doit d'être prise en compte, si l'on veut exprimer correctement la conceptualisation du domaine audiovisuel qui nous intéresse. De plus, nous avons vu qu'il est tout à fait envisageable de faire co-exister les deux points de vue de manière cohérente dans un même SBC, pourvu que l'on y incorpore des connaissances de raisonnement adaptées...

Le dernier point est semblable, mais il se distingue par le fait que les choix de représentation qu'il implique sont directement reliés à un choix applicatif précis. En effet, si la représentation des événements de production et de diffusion qui créent le document audiovisuel est négligée dans notre application de description documentaire, c'est que de notre point de vue leur résultat importe plus que leur déroulement. Ainsi, lorsqu'on voudra exprimer qu'une séquence est tournée en gros plan, on préférera l'exprimer de cette façon même, tel que c'est le cas dans le patron d'utilisation. Il est plus naturel de dire qu'une action a comme valeur de plan la valeur « gros plan », plutôt que de dire qu'elle a joué le rôle de produit d'une activité de tournage faisant partie d'une séquence de production, activité dont le paramètre « valeur de plan » était « gros plan ». Surtout si de tels concepts ne seront jamais utilisés par l'application concernée...

Mais là encore, nous avons montré qu'il est possible, au côté du vocabulaire ontologique et des patrons qui le mobilisent, de spécifier des connaissances de raisonnement qui établissent des règles de passages d'une représentation à une autre. Cette approche permet de respecter la conceptualisation que l'on rencontre dans l'application visée, tout en la rendant compatible avec une vue plus cohérente sur le domaine plus général dont elle relève. *In fine*, cela contribue à une *explicitation rationalisée* de l'engagement ontologique. En quelque sorte, on sait à présent comment l'on parle des entités du domaine, mais on a en plus un aperçu des raisons pour lesquelles on en parle ainsi. Et grâce à ce travail de conception ontologique, l'utilisateur d'une application peut continuer à évoluer dans un contexte avec lequel il est familier.

patrons et interopérabilité sémantique entre SBC. Nous avons défendu l'idée d'une articulation rationnelle entre les conceptualisations des applications et d'autres plus abstraites et plus cohérentes. Cela est en grande partie lié au fait que nous sommes convaincus, comme beaucoup, que l'amélioration de la qualité intrinsèque des ressources ontologiques est nécessaire pour faciliter leur conception – aussi paradoxale cette affirmation puisse paraître au premier abord –

[46] Il faut concéder que ces patrons de conception, et en particulier celui que nous avons présenté, peuvent être d'inspiration cognitive, et à ce titre prendre quelques libertés avec la rigueur métaphysique, si celles-ci correspondent à une réalité cognitive « observable ». Il reste que par rapport à des conceptualisations ancrées dans un domaine les patrons de conceptions restent parés d'une certaine aura d'objectivité.

et leur utilisation. Mais nous ne perdons pas de vue un autre objectif fondamental des approches ontologiques de la représentation de connaissances, et en particulier du web sémantique : celui d'une meilleure ré-utilisation des ressources et des systèmes qui les utilisent.

En effet, si l'on relie le contenu et le fonctionnement d'un SBC dédié à une seule application à des modélisations relativement consensuelles, on augmente les chances que celui-ci aura d'être utilisé par des agents qui ne sont pas immergés dans le contexte applicatif considéré. Et ceci, que ces agents soient des utilisateurs humains ou d'autres systèmes qui utilisent les services du SBC considéré. Car si les conceptualisations de deux systèmes sont conçues en référence explicite à une conceptualisation de haut niveau partagée, ou s'ils font référence à des conceptualisations abstraites pour lesquelles des connaissances d'*alignement* existent déjà, il devient plus facile de les comparer.

De fait, l'utilisation des connaissances de raisonnement dédiées au passage entre une conceptualisation métier et une conceptualisation de haut niveau pourrait faciliter le travail de mécanismes d'alignement automatique d'ontologies, thème en vogue dans le web sémantique à l'heure où ce manuscrit a été écrit[47]. En particulier, alors que la plupart des algorithmes ont pour cible la « simple »création de liens de correspondance entre des concepts provenant d'ontologies différentes, on pourrait envisager une nouvelle génération d'outils qui renverraient un « différentiel de connaissances » entre les structures dans lesquelles ces concepts sont habituellement employées. Ces connaissances, reposant sur des règles de raisonnement riches et précises, car conçues manuellement lors de la construction de l'ontologie, pourraient être utiles pour évaluer la compatibilité entre deux concepts, autant que peuvent l'être des définitions formelles de ces concepts dans un langage comme OWL. Il est cependant évident qu'une telle piste de recherche reste largement prospective, et qu'elle n'est évoquée ici que pour donner un indice de l'intérêt que peut revêtir un alignement correct – même si cela implique qu'il soit manuel – entre des connaissances applicatives et des connaissances de haut niveau.

problèmes de mise en œuvre des connaissances d'alignement. En effet, si les connaissances de raisonnement permettant l'alignement entre les patrons de conception et les patrons d'utilisation sont utiles du point de vue de la construction et de la compréhension des ontologies, leur utilisation concrète dans un SBC peut entraver le bon fonctionnement de celui-ci.

Des travaux partageant nos préoccupations [RZS+99][48] font effectivement mention des problèmes de calculabilité et d'expressivité relatifs à de telles solutions. Et les représentations intermédiaires introduites ne sont pas gérées par les mécanismes de raisonnement ontologique standards, mais par des outils auxiliaires, qui par exemple font des traitements sur les représentation avant de les introduire dans la base de connaissance. Ce choix est en partie dû à la décision fondamentale d'exclure ces « connaissances d'encodage » des ontologies – on retrouve là la différence entre les deux classes de connaissances ontologiques que nous avons évoquée en page 192 – mais aussi aux capacités parfois restreintes des systèmes d'inférence utilisés. Comme nous l'avons également indiqué dans cette section, les connaissances de raisonnement dont nous recommandons l'utilisation dans les SBC d'indexation induisent déjà un niveau de complexité important[49]. Appliquées sur des bases de connaissances importantes, en conjonction avec des connaissances d'alignement avec les structures et notions de haut niveau, elles pourraient tout bonnement rendre le système non opérationnel.

[47]Voir par exemple la campagne d'évaluation d'alignements ontologiques sur le site `http://oaei.inrialpes.fr/`, ou bien l'atelier *Integrating Ontologies* de la conférence K-CAP 2005.

[48]Dans cette approche, il s'agit d'introduire des représentations intermédiaires entre l'utilisateur et la connaissance exprimée formellement, ce qui rapproche ces travaux de ce que nous avons proposé.

[49]En particulier par rapport à ce que les outils standard du web sémantique peuvent proposer

Si nous-mêmes considérons que les patrons et leur articulation avec des connaissances plus abstraites font bien partie de la spécification d'une conceptualisation *située*, nous ne devons cependant pas négliger ce problème de l'opérationnalisation des connaissances de raisonnement. A un travail qui a validé une hypothèse relative à l'utilisation des ontologies – les patrons d'indexation – puis détaillé des éléments de conception et d'exploitation de ces connaissances – le couplage entre patron de conception et d'utilisation – il conviendrait d'ajouter des tests opérationnels à plus grande échelle. Ces tests pourraient ensuite déboucher sur des stratégies adaptées d'implémentation, comme le choix de « saturer » une base de connaissances lorsqu'on y ajoute de nouvelles informations, pour soulager le travail du moteur d'inférence lors de la recherche.

vers des bibliothèques de patrons et d'utilisation ? Les démarches d'utilisation de patron de conception s'inspirent des solutions de *design patterns* de l'ingénierie logicielle [GHJV95]. Comme eux, ils s'attachent à proposer des structures qui constituent des solutions à des problèmes précis – en l'occurrence, des difficultés de *modélisation*. Mais ces approches, la nôtre y compris, gagneraient peut-être à s'inspirer plus encore que cela n'est fait à présent de certaines propositions méthodologiques de ce domaine. En particulier, partant du principe que les systèmes à construire répondent à des problèmes complexes, certains chercheurs proposent de constituer des *catalogues* de *design patterns* où ceux-ci sont liés les uns aux autres. Cela permet de mieux cerner les solutions existantes lors de la recherche d'un patron, ou de mieux justifier la création d'un nouveau patron, en faisant explicitement référence à ceux qu'il adapte en vue de résoudre un problème que l'on juge nouveau.

Si l'on était en mesure d'adapter convenablement de telles approches au cas ontologique, où les connaissances sont aussi difficiles à appréhender, à mettre en relation ou à ré-utiliser, on pourrait améliorer les démarches :

– de création de nouveaux patrons de conception ontologiques en référence à des patrons de conception existants ;
– de création de nouveaux patrons d'utilisation ontologiques en référence à des patrons d'utilisation existants ;
– de sélection de patrons de conception ontologiques pour la création des patrons d'utilisation ontologiques.

Dans le domaine des patrons logiciels, des travaux comme ceux de [RGSF99] (cité par [Gza00]) font la distinction entre quatre relations particulières.

alternative. Un patron A est une alternative d'un patron B si A a le même problème que B mais propose une solution différente.

raffine. Un patron A raffine un patron B si le problème résolu par A est une spécialisation (un cas particulier) de celui résolu par B. B peut résoudre les problèmes résolus par A. On dit aussi que B *étend* A.

requiert. Un patron A requiert un patron B si l'application de B est un pré-requis de l'application de A.

utilise. Un patron A utilise un patron B si une partie des problèmes posés par A peut être résolue en partie ou complètement par B. Ainsi la solution-démarche de A est exprimée en utilisant le patron B.

Un programme de recherche en matière de constitution de bibliothèques de patrons de conception et d'utilisation ontologiques impliquerait tout d'abord d'étudier la signification que peuvent prendre de telles relations, formulées en des termes métiers d'ingénierie logicielle, dans un cadre ontologique. Par exemple, on peut supposer que la relation d'« affinage » entre un patron ontolo-

gique A et un patron ontologique B correspondrait à une spécialisation[50] entre A et B – le patron le plus « raffiné » spécialisant celui qui l'est le moins. Ensuite, il faudrait voir si de nouvelles relations possiblement spécifiques au cadre ontologique peuvent être introduites. Si la relation d'affinage peut se traduire par une implication logique entre deux patrons, cette implication peut de son côté correspondre à un lien plus général : ainsi, la relation d'utilisation semble également pouvoir s'exprimer en termes d'implication, puisqu'elle induit l'inclusion d'un patron dans un autre. . . . Il resterait donc à déterminer ce à quoi peut correspondre *exactement* la relation de spécialisation entre patrons ontologiques. Enfin, il serait intéressant de proposer des outils permettant d'assister la gestion de telles bibliothèques, en particulier pour l'utilisateur qui cherche à créer de nouveaux patrons : interfaces, formats de publication, mécanismes de partage. . .

5.4 Conclusion

Ce chapitre vient donc de restituer leur contexte expérimental aux multiples exemples dont nous avons émaillé le début de cette thèse. En effet, les réflexions méthodogiques que nous avons présentées dans les chapitres précédents ont bien été mises en œuvre dans des exercices concrets et significatifs de modélisation. Ces efforts d'application ont ensuite, en retour, contribué au mûrissement de la dite méthodologie. Il faut d'ailleurs noter qu'un tel retour n'a été possible et légitime que parce que nos exercices sont loins de n'avoir constitué que de simples jouets. En effet, sur les cinq expérimentations[51] que nous avons présentées, une seule n'a pas été explicitement réalisée en vue de l'usage d'indexation auquel nous destinions nos réflexions. Et, même si elles sont loin de faire de ceux-ci les contributions les plus significatives de l'ingénierie ontologique, la taille ainsi que la richesse des modèles à la réalisation desquels nous avons participé n'ont strictement rien de honteux.

Il nous faut néanmoins reconnaître que nous avons manqué d'utilisateurs « réels » – à l'exception notable du cas d'Opales. Nos expérimentations correspondent bien à des besoins et des usages observables, mais en l'absence de validation finale par l'utilisateur – ou d'un concepteur différent de nous-même – la qualité de notre analyse peut être prise en défaut. Nous pouvons cependant nuancer cette remarque par le fait que ces expériences bénéficient d'un certain « effet de nombre ». Du fait de l'organisation de la recherche sur les ontologies à l'INA, notre travail d'acquisition de connaissances a été en effet toujours mené en collaboration avec d'autres chercheurs, et s'est systématiquement appuyé – que nous ayons participé ou non à ces efforts de justification – sur des sources attestées dans les domaines visés : guides de bonnes pratiques, textes relatifs au thèmes abordés, élicitation de connaissances auprès d'experts des domaines d'application visés. . .

Et pourtant, comme l'a montré la seconde partie de ce chapitre, notre travail de recherche serait loin d'être terminé. La mise en œuvre de nos propositions a directement soulevé de nombreux problèmes, auxquels nous n'avons pas forcément pu apporter de réponses. Nous pourrions nous défausser en rappelant que nombre d'entre eux n'étaient pas vraiment de notre ressort, tel que celui de la création d'un outillage pouvant réaliser des calculs d'inférence complexes sur une grande échelle. Il reste que ces problèmes sont inhérents à l'approche que nous proposons, et que le lecteur devra garder ce fait à l'esprit lors de la lecture de la conclusion de ce manuscrit, qui va aborder des questions plus générales concernant l'élargissement de la portée de nos travaux.

[50] Au sens d'une implication logique entre leurs interprétations formelles, comme dans le cas des gc.

[51] Il ne nous apparaît pas illégitime de considérer notre participation au projet Opales comme constituant deux expérimentations ontologiques à part entière. . .

Nous espérons cependant que l'effort d'exemplification et de discussion auquel nous nous sommes livré ici pourra être utile à ceux qui voudraient ré-employer nos hypothèses, à fins d'utilisation directe ou de recherche nouvelle.

Conclusion

Adéquation de nos solutions aux questions posées

Initialement, le sujet de ce manuscrit aurait dû être « *Capture et formalisation des contraintes sémantiques : utilisation pour la conception des ontologies* ». A l'arrivée, le lecteur pourrait très bien faire observer que, d'une certaine façon, ce sujet a été traité dans ce manuscrit. Tout au long de nos travaux, nous nous sommes effectivement intéressé à la façon dont on pouvait rendre compte, dans un cadre formel, de significations et d'usages relevant d'un contexte sémantique non formel. Ce contexte interprétatif « métier » est défini par les interprétations que l'on rencontre dans le domaine applicatif visé par un système d'information particulier. Et on peut considérer que ces interprétations apportent justement des contraintes qui *via* leur transposition connaissances formelles – expressions définitoires, règles d'inférence – devraient idéalement s'appliquer à l'interprétation formelle du vocabulaire de description apporté par une ontologie.

Cette thèse propose donc un cadre de conception d'ontologies qui permet une utilisation rationnelle et relativement efficace de ces ressources dans les processus d'indexation et de recherche. On retrouve au cours de nos discussions des critères employés traditionnellement pour définir ou évaluer une ontologie : pertinence, consensus, partage, interopérabilité, etc. Mais la mise en pratique de ces notions abstraites est évidemment indissociable d'une réflexion complète sur l'utilisation de ces ontologies dans un contexte concret, et, partant, sur la manière d'acquérir un niveau de pertinence satisfaisant par rapport aux pratiques observables. Le déroulement de notre thèse a donc fait que nous nous sommes intéressé à une problématique plus large que la simple *capture* de contraintes interprétationnelles : du fait de l'ancrage de nos travaux dans des projets concrets d'utilisation d'ontologies, ce dernier aspect a pris une part plus importante au fur et à mesure de nos réflexions... Finalement, on peut constater que ce que cette thèse apporte, c'est un ensemble de points méthodologiques qui permettent de situer correctement les activités de conception d'ontologies, mais qui concernent également leur utilisation pour la création de bases de connaissances qui puissent être correctement interprétées par les utilisateurs humains et par les systèmes formels eux-mêmes. Plus important est le fait que ces points ne constituent pas des idées indépendantes, mais un *cadre* à qui l'on s'est efforcé de donner une réelle cohérence. La conception des ontologies se fait en étant réellement conscient de l'usage des primitives en cours de définition, et, réciproquement, l'utilisation de ces ontologies pendant l'indexation et la recherche se veut exploiter au mieux les connaissances qui ont été spécifiées lors de la conception.

Ces résultats semblent en conformité avec ce que l'on attend généralement d'une convention CIFRE. Car si à présent on peut relier les préoccupations scientifiques à des applications pratiques, en observant comment on a pu limiter la complexité induite par le choix de techniques élaborées, on sait aussi comment directement bénéficier de certains des efforts les plus théoriques – comme les ontologies de haut niveau ou les patrons de conception – pour améliorer la qualité intrinsèque des ressources construites pour ces applications.

Par exemple, le recours à des connaissances de raisonnement formelles peut tout d'abord être

interprété en termes de légitimité métier. En compensant les variations descriptives autorisées autour du contenu de référence du graphe patron d'indexation[52], les connaissances de raisonnement permettent de prendre en compte la variabilité inter- ou intra-indexeurs (cf. page 35) dans les processus de recherche assistée. En termes de performances de système de recherche d'informations, cela permet de gagner en rappel sans perdre en précision, puisque la stratégie d'élargissement de l'ensemble des résultats positifs correspond à une connaissance dont on est sûr de la pertinence.

Ensuite, ce même recours à des connaissances de raisonnement permet, en articulant les structures construites dans le cadre de l'application avec des constructions de haut niveau – les patrons de conception – de légitimer celles-ci d'un point de vue théorique. Ceci augmente leur potentiel de partageabilité, et, même si dans cette thèse cela n'est pas notre objectif immédiat, leur ré-utilisabilité dans d'autres contextes applicatifs. On peut considérer que ces connaissances permettent de compenser une variabilité inter-applications, ce qui est utile pour atteindre les objectifs – jusque-là énoncés de façon plutôt théorique – d'initiatives de partage de connaissances à grande échelle.

Il faut noter que ce genre de considérations rapproche en un certain sens nos vues de celles de l'initiative du web sémantique. Cette initiative, au lieu de proposer des paradigmes en rupture totale avec les solutions techniques ou méthodologiques précédentes, vise en effet davantage un aménagement de celles-ci pour des contextes d'application bien particuliers.

On pourra faire remarquer que nous n'avons pas instrumenté nous-mêmes l'intégralité de la chaîne de conception d'ontologies puis d'indexation et de recherche de documents audiovisuels que nous proposions. Certains outils sont en effet déjà disponibles, ou ont été développés par d'autres que nous dans le projet dans lequel nous étions impliqué au début de cette thèse. Nous avons cependant créé un éditeur d'ontologie approprié au moment où il le fallait, et discuté des conditions de réutilisation des outils et techniques proposés dans notre champ, *via* l'évocation des fonctionnalités qui sont indipensables à la mise en œuvre de nos propositions. Ainsi, en matière de conception formelle d'ontologie et d'utilisation de mécanismes de raisonnement, il faut évidemment que l'expressivité autorisée par le moteur d'inférences soit importante, notamment en ce qui concerne l'utilisation des relations entre entités. Et si nous n'avons pas détaillé les modalités de prise en compte des patrons d'indexation lors de la création des descriptions, c'est parce que les fonctions requises – telles que la possibilité de stocker un graphe, et le copier dans une annotation, comme cela a été fait dans Opales – ne semblaient pas complexes. Par contre, un cadre technique permettant la création des patrons d'indexation, la gestion des liens entre ces patrons, mais aussi celle des liens entre patrons de conception et patrons d'utilisation aurait été bienvenu. Surtout s'il avait permis ces activités en liaison avec la spécification des autres types de ressources ontologiques auxquelles ces patrons sont liés, comme les connaissances de raisonnement.

Plus gênant est le constat que tous les obstacles techniques en rapport avec ces fonctionnalités ne sont pas levés. La mise en œuvre de nos solutions reste évidemment dépendante des capacités des langages de représentation proposés, et des performances des systèmes qui réalisent les raisonnements demandés. Comme on a pu le voir dans le chapitre 5, le cadre que nous proposons exige des connaissances relativement riches, ce qui augmente la complexité que des moteurs d'inférence adaptés auront à gérer. Et s'il n'a jamais été dans nos objectifs de réaliser des outils permettant de résoudre ce problème, force est de reconnaître que davantage de tests quantitatifs

[52]On peut remarquer que cette remarque reste valable même dans le cas où l'on n'utiliserait pas la structure pivot que constitue ce graphe patron.

auraient été nécessaires : ontologies encore plus volumineuses en termes de connaissances de raisonnement complexes, bases d'index correspondant à l'activité d'une communauté précise, mais sur une durée bien plus longue que celle qui a pu être simulée dans nos expérimentations... Et, comme toute méthode fondamentalement différente de celles qui la précèdent, l'indexation ontologique devra toujours faire l'objet d'un apprentissage, et impliquera nécessairement un certain niveau d'expertise, aussi bien guidé l'utilisateur soit-il. Il reste que les hypothèses que nous avons avancées ont été testées, et avec succès, pour des scénarios réalistes dans le contexte des documents audiovisuels. Ceci montre qu'au moins pour les personnes interagissant avec le système – concepteurs, indexeurs, utilisateurs effectuant des recherches – notre approche est pertinente, et praticable.

Est-il possible de généraliser nos propositions ?

Nous avons déjà discuté, tout au long de ce manuscrit, et surtout dans le chapitre 5, des problèmes et perspectives soulevés par quelques points bien précis de notre cadre méthodologique et technique. Dans cette conclusion, nous ne reviendrons pas sur ceux-ci, mais préférons nous concentrer sur la *portée* de ces hypothèses. En effet, nous nous sommes depuis le début de nos travaux principalement intéressé au cadre bien particulier d'un indexation riche – et manuelle – de documents audiovisuels. Il est légitime de se demander si nos réflexions et solutions peuvent sortir de ce cadre et être transposées dans d'autres contextes applicatifs. Nombre des outils, techniques ou points méthodologiques abordés dans cette thèse semblent en effet avoir une portée plus générale.

Ainsi, la méthodologie de conception d'ontologies utilisant la normalisation sémantique est utile parce qu'elle permet de prendre en compte, lors de la construction des hiérarchies de concepts et de relations, la signification métier des descripteurs employés dans les index – et aussi de restituer convenablement celle-ci quand on accède à l'ontologie lors de la création des index. Il est naturel d'affirmer qu'une telle solution sera d'un intérêt majeur dès que des sujets humains seront en contact avec les entités de l'ontologie. Si une telle confrontation n'est pas acquise en ce qui concerne l'utilisateur final[53], on peut imaginer que ces informations seront toujours utiles au concepteur des ontologies, même – et peut-être surtout – s'il doit être assisté d'outils de construction semi-automatique. Il faut cependant mettre un bémol à cette affirmation : l'application de cette méthode, tout à fait apte à désambiguiser le sens des concepts dans un contexte applicatif précis, peut se révéler contre-productive dans des cas où l'on recherche justement une modélisation tenant compte de certaines ambiguités et de rapprochements que l'on n'autoriserait pas dans le genre d'ontologies que nous souhaitons utiliser. Par exemple, WORDNET est un répertoire terminologique qui essaie de représenter des relations d'hyperonymie, de synonymie, etc... dans le domaine de la langue générale [Fel98]. Sa structure est donc relativement peu rigoureuse, comme l'ont montré les travaux d'Oltramari et de ses collègues [OGGM02]. Pourtant, cet artefact est bel et bien catalogué comme une ontologie – aussi contestable puisse être ce rattachement d'un point de vue théorique – et il est reconnu comme utile du fait même du flou

[53]Encore que la justification des résultats d'un système, sujet relativement présent dans le cadre de l'intelligence artificielle et du raisonnement automatique en particulier, puisse, au fil des recherches, se révéler être d'un intérêt crucial pour les types de SBC actuellement en vogue. Tim Berners-Lee, dans son désormais célèbre « layer cake » qui présente les différentes couches de services offerts par un web sémantique, place en effet les « preuves » et la « confiance » au sommet de l'échelle [BHL01]. Mais les recherches dans le domaine n'ont pas encore montré la mesure dans laquelle ces fonctionnalités impliquaient la présentation des entités de l'ontologie à l'utilisateur final de systèmes.

des relations qu'il introduit[54]. En première instance, on peut donc affirmer que la méthodologie de normalisation différentielle – et, par extension, notre éditeur qui la met en œuvre – est utile dès qu'un agent humain a besoin d'accéder au contenu d'une ontologie s'inscrivant dans un cadre applicatif précis. Ce qui concerne évidemment une quantité non nulle d'applications à base de connaissances.

Ensuite, le spectre des situations où on peut avoir recours aux patrons d'indexation peut également être élargi, dès le moment où l'on cesse de les considérer comme des structures uniquement employables par des indexeurs humains, mais comme des ressources pouvant guider tout *agent* susceptible de créer des représentations à la structure et au contenu récurrents. Par exemple, on a vu que les applications d'extraction d'information ontologiques à partir de textes pouvaient avoir besoin de telles constructions pour orienter leur travail [ZC94, Le 03]. Des approches d'extraction d'informations parcourant des bases de données pour associer les valeurs rencontrées à des entités du domaine et en déduire des relations entre celles-ci – voir par exemple [HS03a] – peuvent également avoir intérêt à déterminer explicitement des formes correspondant aux descriptions typiques qu'elles sont supposées extraire des données structurées préexistantes.

L'approche de conception d'ontologies par patrons de conception est également relativement généralisable. Dans la mesure où l'ontologie à concevoir doit contenir plus qu'une simple hiérarchie de concepts sans relations conceptuelles pour lier explicitement ceux-ci, on peut être intéressé par la réutilisation et l'adaptation de structures relationnelles ayant fait l'objet d'un effort de spécification important. Tout dépendra ensuite de l'existence d'un patron correspondant à un usage générique proche de celui qui est envisagé[55], et de la possibilité de l'adapter, comme nous l'avons fait, aux contenus réellement rencontrés dans l'application.

Finalement, c'est peut-être l'importance accordée à une forme relativement riche de raisonnement qui distingue le plus notre cadre applicatif de ceux que l'on rencontre traditionnellement, surtout dans le domaine du web sémantique. De nombreuses applications, concentrées sur la seule création des descriptions conceptuelles et l'accès à leur contenu sans plus de formalités, peuvent ne pas demander de raisonnement du tout. Et beaucoup d'autres, préférant faire le choix de la simplicité – mais aussi des performances de leurs systèmes – et n'utiliseront que des formes élémentaires d'inférence, comme celle du raisonnement hiérarchique que l'on peut déjà trouver dans certains systèmes à base de thesauri. Cependant la situation évolue relativement rapidement, et avec la popularisation de langages suffisamment expressifs comme OWL la communauté redécouvre les joies du raisonnement formel complexe comme élément fondamental d'un système à base de connaissances. Et démontre que les solutions actuelles du web sémantique ne sont pas suffisamment complètes, en particulier parce qu'elles n'autorisent pas assez de fonctionnalités de raisonnement exploitant correctement les relations. D'où l'engouement actuel pour les langages et outils de représentations de règles plus génériques[56].

Bien que l'indexation du document audiovisuel ait fourni le cadre de référence de nos travaux, on peut donc envisager de réutiliser nos propositions dans d'autres contextes. Il semble tout d'abord extrêmement naturel d'aller vers l'indexation de documents non audiovisuels. On

[54] WORDNET peut par exemple être utilisé comme connaissance de contexte dans des processus d'*alignement* d'ontologies.

[55] Mais on peut considérer que celui proposé dans [GM03] pourra déjà être réutilisé dans nombre d'applications, son usage descriptif le plaçant au cœur des activités traditionnelles de représentation.

[56] On peut consulter par exemple les articles [BTW01, Wag03] et [HPB+04], mais aussi des initiatives comme le réseau d'excellence européen REWERSE, `http://rewerse.net/`. On remarquera que tout cela est toujours évidemment conforme à la vision du visionnaire Tim Berners-Lee, qui avait astucieusement songé à insérer un item « rules » dans son gâteau.

perdrait certaines des ressources que nous avons pu concevoir pour l'indexation audiovisuelle – notamment l'ontologie de l'audiovisuel développée avec Raphaël Troncy [ICG^{+}04], mais la méthodologie est toujours applicable. D'autres contextes sont envisageables ; le tout est d'être confronté à une application qui demande des descriptions riches, qui offre les moyens de les obtenir – même si les structures spécialisant les patrons d'utilisation sont obtenues par extraction d'information automatique – et qui requiert des traitements inférentiels élaborés à partir du contenu – et en particulier du contenu relationnel – de ces descriptions .

Bibliographie

[Abr95] Abraham M. : *Analyse Sémantico-Cognitive des verbes de mouvement et d'activité, Contribution méthodologique à la construction d'un dictionnaire informatique des verbes*. Thèse de doctorat, EHESS, 1995.

[ACFG01] Arpirez J.C., Corcho O., Fernández-López M. et Gómez-Pérez A. : WebODE : a Workbench for Ontological Engineering. *In First International Conference on Knowledge Capture (K-CAP'01)*, Victoria, Canada, 2001.

[AFN87] AFNOR : Vocabulaire de la documentation. Rapport technique, AFNOR, 1987.

[AFN93] AFNOR : Information et documentation - Principes généraux pour l'indexation des documents, NF Z47-102. Rapport technique, AFNOR, 1993.

[AFS00] Amann B., Fundulaki I. et Scholl M. : Integrating Ontologies and Thesauri for RDF Schema Creation and Metadata Querying. *International Journal of Digital Libraries*, 3(3), 2000.

[AHAS03] Aleman-Meza B., Halaschek C., Arpinar I. B. et Sheth A. : Context-Aware Semantic Association Ranking. *In First International Workshop on Semantic Web and Databases (SWDB'03)*, Berlin, Germany, 2003.

[ALR96] Aussenac-Gilles N., Laublet P. et Reynaud C. : L'acquisition des connaissances : une composante à part entière de l'informatique du futur. *In* Aussenac-Gilles N., Laublet P. et Reynaud C., éditeurs : *Tendances actuelles en acquisition et modélisation des connaissances*. Cépadues, 1996.

[AMO+03] Angele J., Moench E., Oppermann H., Staab S. et Wenke D. : Ontology-Based Query and Answering in Chemistry : OntoNova@Project Halo. *In International Semantic Web Conference (ISWC 2003)*, Sanibel Island, Florida, USA, 2003.

[ANN03] Alhulou R., Napoli A. et Nauer E. : Une mesure de similarité sémantique pour raisonner sur des documents. *In 3èmes Journées Nationales sur les Modèles de Raisonnement (JNMR'03)*, Paris, 2003.

[Arp00] Arpirez J.C. : WebODE 1.0 User's Manual. `http://delicias.dia.fi.upm.es/webODE/`, 2000.

[AS92] Amann B. et Scholl M. : Gram : A Graph Data Model and Query Language. *In Proceedings International Workshop on Semantic Web Foundations and Application Technologies*, Milano, Italy, 1992.

[Auf00] Auffret G. : *Structuration de documents audiovisuels et publication électronique*. Thèse de doctorat, Université de Technologie de Compiègne, 2000.

[Ba00] Bechhofer S. et alii : OIL whitepaper. `http://www.ontoknowledge.org/oil/index.shtml`, 2000.

[Bac96] Bachimont B. : *Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser*. Thèse de doctorat, Ecole Polytechnique, 1996.

[Bac98] Bachimont B. : Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques. *Document numérique*, 2(3), 1998.

[Bac00] Bachimont B. : Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. *In* Charlet J., Zacklad M., Kassel G. et Bourigault D., éditeurs : *Ingénierie des Connaissances : Evolutions récentes et nouveaux défis*. Eyrolles, 2000.

[Bac01] Bachimont B. : Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle. *In Actes de la conférence en Ingénierie des Connaissances : IC'2001*, Grenoble, France, 2001.

[Bac04a] Bachimont B. : *Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle*. Habilitation à diriger des recherches, Université de Technologie de Compiègne, 2004.

[BAC04b] Bourigault D., Aussenac-Gilles N. et Charlet J. : Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'intelligence artificielle*, 18(1), 2004.

[Bag03] Baget J.-F. : Homomorphismes d'hypergraphes pour la subsomption en RDF. *In 3èmes Journées Nationales sur les Modèles de Raisonnement (JNMR'03)*, Paris, 2003.

[BBCZ95] Bouaud J., Bachimont B., Charlet J. et Zweigenbaum P. : Methodological principles for structuring an ontology. *In IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Montréal, 1995.

[BCKN01] Barry C., Cormier C., Kassel G. et Nobécourt J. : Evaluation de langages opérationnels de représentation d'ontologies. *In Actes de la conférence en Ingénierie des Connaissances : IC'2001*, Grenoble, France, 2001.

[BCM+03] Baader F., Calvanese D., McGuinness D., Nardi D. et Patel-Schneider P.F. : *The Description Logic Handbook : Theory, Implementation and Applications.* Cambridge University Press, 2003.

[BFGG98] Blázquez M., Fernández M., García-Pinar J.M. et Gómez-Pérez A. : Building Ontologies at the Knowledge Level using the Ontology Design Environment. *In 11th Banff Knowledge Acquisition Workshop (KAW'98)*, Banff, Canada, 1998.

[BHGS01] Bechhofer S., Horrocks I., Goble C. et Stevens R. : OilEd : a Reason-able Ontology Editor for the Semantic Web. *In Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence*, Vienna, 2001.

[BHL01] Berners-Lee T., Hendler J. et Lassila O. : The Semantic Web. *Scientific American*, 284(5), 2001.

[BIT02] Bruno Bachimont, Antoine Isaac et Raphaël Troncy : Semantic Commitment for Designing Ontologies : A Proposal. *In* A. Gomez-Pérez et V. R. Benjamins, éditeurs : *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, volume Lecture Notes in Artificial Intelligence, vol 2473, Sigüenza, Spain, 2002. Springer Verlag.

[BKM03] Bruaux S., Kassel G. et Morel G. : Étude critique de la méthode CommonKADS, application au calage de codes de calcul. *In Actes de la Conférence en Ingénierie des Connaissances IC'2003*, Laval, 2003.

[BKv02] Broekstra J., Kampman A. et van Harmelen F. : Sesame : a Generic Architecture for Storing and Querying RDF and RDF Schema. *In International Semantic Web Conference (ISWC 2002)*, Sardinia, Italia, 2002.

[BLGP03] Bayerl P. S., Lüngen H, Gut U. et Paul K. I. : Methodology for Reliable Schema Development and Evaluation of Manual Annotations. *In Proceedings of the Second International Conference on Knowledge Capture K-CAP 2003, Workshop on Knowledge Markup and Semantic Annotation (Semannot 2003)*, Sanibel Island, Florida, 2003.

[Blo04] Blocks D. : *A qualitative study of thesaurus integration for end-user searching.* Thèse de doctorat, Univerity of Glamorgan, 2004.

[BLS+00] Benslimane D., Leclerq E., Savonnet M., Terrasse M.-N. et Yétongnon K. : On the definition of generic multi-layered ontologies for urban applications. *Computers, Environments and Urban Systems*, 24(3), 2000.

[BMP+91] Brachman R. J, McGuiness D. L., Patel-Schneider P. F., Resnik L. A. et Borgida A. : Living with CLASSIC : When and how to use a KL-ONE-like language. *In* Sowa J. F., éditeur : *Principles of Semantic Networks.* Morgan Kaufmann Publishers, San Mateo, California, 1991.

[Bra79] Brachman R. J. : On the Epistemological Status of Semantic Networks. *In* Findler N. V., éditeur : *Associative Networks : Representation and Use of Knowledge by Computers.* Academic Press, 1979.

[BTW01] Boley H., Tabet S. et Wagner G. : Design Rationale of RuleML : A Markup Language for Semantic Web Rules. *In Proceedings of the International Semantic Web Working Symposium (SWWS)*, Stanford University, California, USA, 2001.

[Car00] Carrive J. : *Classification des séquences audiovisuelles.* Thèse de doctorat, Université Paris VI, 2000.

[Cha03] Charlet J. : *L'ingénierie des connaissances. Développements, résultats et perspectives pour la gestion des connaissances médicales.* Habilitation à diriger des recherches, Université Pierre et Marie Curie, 2003.

[CM01] Ceusters W. et Martens P. : LinkFactory : an Advanced Formal Ontology Management System. *In Interactive Tools for Knowledge Capture Workshop, KCAP-2001*, Victoria, Canada, 2001.

[CMF96] Chiaramella Y., Mulheim P. et Fourel F. : fermi (Formalization and Experimentation in the Retrieval of Multimedia Information) Report. A model for multimedia information retrieval. Rapport technique, CLIPS-IMAG, 1996.

[CTHD00] Clark P., Thompson J., Holmback H. et Duncan L. : Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search. *In Proceedings of the 12th Innovative Applications of AI Conference (IAAI'00)*, Austin, Texas, USA, 2000.

[CTP00] Clark P., Thompson J. et Porter B. : Knowledge Patterns. *In Proceedings of the 7th Conference on Principles of Knowledge Representation and Reasoning (KR2000)*, Breckenridge, Colorado, USA, 2000.

[CZKB00] Charlet J., Zacklad M., Kassel G. et Bourigault D. : *Ingénierie des Connaissances : Evolutions récentes et nouveaux défis.* Eyrolles, 2000.

[Dau94] Dauzats M. : *Le thesaurus de l'image : étude des langages documentaires pour l'audiovisuel.* ADBS Editions, Paris, 1994.

[DB00] DECHILLY T. et BACHIMONT B. : Une ontologie pour éditer des schémas de description audiovisuels, extension pour l'inférence sur les descriptions. *In Actes de la conférence IC'2000 : Journées francophones d'Ingénierie des Connaissances*, Toulouse, 2000.

[DDH+04] DASMAHAPATRA S., DUPPLAW D., HU B., LEWIS H., LEWIS P. et SHADBOLT N. : Facilitating multi-disciplinary knowledge-based support for breast cancer screening. *International Journal of Healthcare Technology and Management*, 2004.

[Den04] DENNY M. : Ontology Tools Survey, Revisited. `http://www.xml.com/pub/a/2004/07/14/onto.html`, July 2004.

[Des87] DESCLÉS J.-P. : Réseaux sémantiques : La nature logique et linguistique des relateurs. *Langages*, 87, 1987.

[Des90] DESCLÉS J.-P. : *Languages applicatifs, langages naturels et cognition.* Hermès, Paris, 1990.

[Dev99] DEVEDZIC V. : Ontologies : Borrowing from Software Patterns. *ACM Intelligence Magazine*, 10(3), 1999.

[DHHS01] DEGEN W., HELLER B., HERRE H. et SMITH B. : GOL : Towards an Axiomatized Upper-Level Ontology. *In Formal Ontology and Information Systems (FOIS-2001)*, Ogunquit, Maine, USA, 2001.

[Dji00] DJIOUA B. : *Modélisation Informatique d'une base de connaissances lexicales (DiSSC), Réseaux polysémiques et Schèmes Sémantico-Cognitifs.* Thèse de doctorat, Université de Paris IV – Sorbonne, 2000.

[Dom98] DOMINGUE J. : Tadzebao and WebOnto : Discussing, Browsing, and Editing Ontologies on the Web. *In 11th Banff Knowledge Acquisition Workshop (KAW'98)*, Banff, Canada, 1998.

[DSW+99] DUINEVELD A. J., STOTER R., WEIDEN M. R., KENEPA B. et BENJAMINS V. R. : Wondertools ? A comparative study of ontological engineering tools. *In 12th Workshop on Knowledge Acquisition, Modeling and Management (KAW'99)*, Banff, Canada, 1999.

[ES03] ENSER P. et SANDOM C. : Towards a Comprehensive Survey of the Semantic Gap in Visual Image Retrieval. *In International Conference on Image and Video Retrieval (CIVR 2003)*, Urbana, Italy, 2003.

[FAD+00] FENSEL D., ANGELE J., DECKER S., ERDMANN M., SCHNURR H.-P., STUDER R. et WITT A. : Lessons Learned from Applying AI to the Web. *Journal of Cooperative Information Systems*, 9(4), 2000.

[Fel98] FELLBAUM C., éditeur. *WordNet – An Electronic Lexical Database.* MIT Press, 1998.

[FFR97] FARQUHAR A., FIKES R. et RICE J. : The Ontolingua Server : A Tool for Collaborative Ontology Construction. *International Journal of Human-Computer Studies*, 46(6), 1997.

[FG02] FERNANDEZ-LOPEZ M. et GOMEZ-PEREZ A. : The Integration of OntoClean in WebODE. *In EKAW 2002 Workshop on Evaluation of Ontology-based Tools (EON2002)*, Siguenza, Spain, 2002.

[FGJ97] FERNÁNDEZ M., GÓMEZ-PÉREZ A. et JURISTO N. : METHONTOLOGY : From Ontological Art Towards Ontological Engineering. *In AAAI97 Spring Symposium Series on Ontological Engineering*, Stanford, USA, 1997.

[Fil68] Fillmore C. J. : The Case for Case. *In* Bach E. et Harms R., éditeurs : *Universals in Linguistic Theory*. Holt, Rinehart & Winston, 1968.

[Für04] Fürst F. : *Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation*. Thèse de doctorat, Université de Nantes, 2004.

[Fre71] Frege G. : *Écrits logiques et philosophiques*. Éditions du Seuil, 1971.

[FvH+01] Fensel D., van Harmelen F., Horrocks I., McGuiness D. L. et Patel-Schneider P. F. : OIL : An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems*, 16(2), 2001.

[GB04] Gangemi A. et Borgo S., éditeurs. *Workshop on Core Ontologies in Ontology Engineering, 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004)*, Whittlebury Hall, Northamptonshire, UK, 2004. CEUR online Proceedings, `http://ceur-ws.org/Vol-118/`.

[GBvH03] Geurst J., Bocconi S., van Ossenbruggen J. et Hardman L. : Towards Ontology-driven Discourse : from Semantic Graphs to Mulmedia Presentations. *In International Semantic Web Conference (ISWC 2003)*, Sanibel Island, Florida, USA, 2003.

[GCB04] Gangemi A., Catenacci C. et Battaglia M. : Inflammation Ontology Design Pattern : an Exercise in Building a Core Biomedical Ontology with Descriptions and Situations. *In* Pisanelli D.M., éditeur : *Ontolgies in Medicine*. IOS Press, Amsterdam, 2004.

[GCG94] Guarino N., Carrara M. et Giaretta, P. : Formalizing Ontological Commitments. *In National Conference on Artificial Intelligence (AAAI-94)*, Seattle, 1994. Morgan Kaufmann.

[GDGB03] Golbreich C., Dameron O., Gibaud B. et Burgun A. : Comment représenter les ontologies pour un Web Sémantique Médical ? *In Journées Francophones de la Toile (JFT'2003)*, Tours, France, 2003.

[Gen00] Genest D. : *Extension du formalisme des graphes conceptuels pour la recherche d'information*. Thèse de doctorat, Université Montpellier II, 2000.

[GF95] Grüninger M. et Fox M. : Methodology for the Design and Evaluation of Ontologies. *In IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Montréal, 1995.

[GFC04] Gómez-Pérez A., Fernández-López M. et Corcho O. : *Ontological Engineering*. Springer-Verlag, London, 2004.

[GFP02] Guizzardi G., Falbo R. A. et Pereira Filho J.G. : Using objects and Patterns to implement domain ontologies. *Journal of Brazilian Computer Society (JBCS), Special Issue on Software Engineering*, 8(1), 2002.

[GG95] Guarino N. et Giaretta, P. : Ontologies and Knowledge Bases : Towards a Terminological Clarification. *In* Mars N., éditeur : *Towards Very Large Knowledge Bases : Knowledge Building and Knowledge Sharing*. IOS Press, Amsterdam, 1995.

[GGM+02] Gangemi A., Guarino N., Masolo C., Oltramari A. et Schneider L. : Sweetening Ontologies with DOLCE. *In 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2002)*, Siguenza, Spain, 2002.

[GHB96] Carole Goble, C. Haul et Sean Bechhofer : Describing and Classifying Multimedia using the Description Logic GRAIL. *In SPIE Conference on Storage and Retrieval of Still Image and Video IV*, San Jose, 1996.

[GHJV95] Gamma E., Helm R., Johnson R. et Vlissides J. : *Design Patterns : Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1995.

[GHVD03] Grosof B. N., Horrocks I., Volz R. et Decker S. : Description Logic Programs : Combining Logic Programs with Description Logic. *In Twelfth International World Wide Web Conference (WWW2003)*, Budapest, Hungary, 2003.

[GI04] Golbreich C. et Imai A. : Combining SWRL rules and OWL ontologies with Protégé OWL Plugin, Jess, and Racer. *In 7th International Protégé Conference,*, Bethesda, Maryland, 2004.

[GL02] Graves A. et Lalmas M. : Video Retrieval using an MPEG-7 Based Inference Network. *In ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, 2002.

[Góm99] Gómez-Pérez A. : Développements récents en matière de conception, de maintenance et d'utilisation des ontologies. *Terminologies Nouvelles*, 19, 1999.

[GM03] Gangemi A. et Mika P. : Understanding the Semantic Web through Descriptions and Situations. *In International Conference on Ontologies, Databases and Applications of Semantics (ODBASE'03)*, Catania, Italy, 2003.

[GMV99] Guarino N., Masolo C. et Vetere G. : OntoSeek : Content-Based Access to the Web. *IEEE Intelligent Systems*, 14(3), 1999.

[GPS99] Gangemi A., Pisanelli D. M. et Steve G. : An Overview of the ONIONS project : Applying Ontologies to the Integration of Medical Terminologies. *Data and Knowledge Engineering*, 31(2), 1999.

[Gru93] Gruber T. : A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), 1993.

[Gru95] Gruber T. : Towards Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of human-Computer Studies*, 43(5-6), 1995.

[GS93] Gauch S. et Smith J. B. : An Expert System for Automatic Query Reformulation. *Journal of the American Society of Information Scientists*, 44(3), 1993.

[GS98] Genest D. et Salvat E. : A platform allowing typed nested graphs : How CoGITo became CoGITaNT. *In ICCS'98 : 6th International Conference on Conceptual Structures*, Montpellier, France, 1998.

[Gua92] Guarino N. : Concepts, Attributes and Arbitrary Relations : Some linguistic and ontological criteria for structuring knowledge bases. *Data and Knowledge Engineering*, 8(3), 1992.

[Gua95] Guarino N. : The Ontological Level. *In* Casati R., Smith B. et White G., éditeurs : *Philosophy and the Cognitive Sciences*. Philosophia Verlag, Münich, 1995.

[Gua98a] Guarino N. : Formal Ontology and Information Systems. *In* Guarino N., éditeur : *Formal Ontology and Information Systems (FOIS'98)*, Amsterdam, 1998. IOS Press.

[Gua98b] Guarino N. : Some Ontological Principles for Designing Upper Level Lexical Resources. *In Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.

[GW00a] Guarino N. et Welty C. : A formal ontology of properties. *In* Dieng R. et Corby O., éditeurs : *Knowledge Engineering and Knowledge Management : Methods, Models and Tools. 12th International Conference, EKAW2000*. Springer-Verlag, 2000.

[GW00b] Guarino N. et Welty C. : Identity, unity and individuality : towards a formal toolkit for ontological analysis. *In* Werner H., éditeur : *Proceedings of ECAI-2000 : the European Conference on Artificial Intelligence*, Berlin, Germany, 2000. IOS Press.

[GW01] Guarino N. et Welty C. : Identity and subsumption. *In* Green R., Bean C. A. et Hyon Myaeng S., éditeurs : *The Semantics of Relationships : An Interdisciplinary Perspective*. Kluwer, 2001.

[GW02] Guarino N. et Welty C. : Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 2(45), 2002.

[Gza00] Gzara L. : *Les patterns pour les systèmes d'information produit*. Thèse de doctorat, Institut National Polytechnique de Grenoble, 2000.

[HH01] Heflin J. et Hendler J. : A Portrait of the Semantic Web in Action. *IEEE Intelligent Systems*, 16(2), 2001.

[HM03] Haarslev V. et Möller R. : Racer : A Core Inference Engine for the Semantic Web. *In 2nd International Workshop on Evaluation of Ontology-based Tools (EON2003) at ISWC2003*, Sanibel Island, Florida, 2003.

[HNM02] Hou C.-S. J., Noy N. F. et Musen M. A. : A Template-Based Approach Toward Acquisition of Logical Sentences. *In Intelligent Information Processing 2002, World Computer Congress*, Montréal, Canada, 2002.

[Hor98] Horrocks I. : Using an Expressive Description Logic. FaCT or Fiction ? *In 6th International Conference on Principles of Knowledge Representation and Reasoning (KR'98)*, Trento, Italy, 1998.

[HP04] Horrocks I. et Patel-Schneider P. F. : A Proposal for an OWL Rules Language. *In Proceedings of the Thirteenth International World Wide Web Conference (WWW2004)*, New York City, 2004. ACM Press.

[HPB+04] Horrocks I., Patel-Schneider P. F., Boley H., Tabet S., Grosof B. et Dean M. : SWRL : A Semantic Web Rule Language Combining owl and RuleML. online, may 2004.

[HPv03] Horrocks I., Patel-Schneider P. F. et van Harmelen F. : From SHIQ and RDF to OWL : The Making of a Web Ontology Language. *Journal of Web Semantics*, 1(1), 2003.

[HS03a] Handschuh S. et Staab S. : Annotation of the Shallow and the Deep Web. *In* Handschuh S. et Staab S., éditeurs : *Annotation for the Semantic Web*. IOS Press, Amsterdam, 2003.

[HS03b] Hu B. et Shadbolt N. : Visualising a DL Knowledge Base with DeLogViz. *In Poster Session of the 2003 International Workshop on Description Logics (DL 2003)*, Rome, Italy, 2003.

[HSV04] Hyvönen E., Saarela S. et Viljanen K. : Application of Ontology Techniques to View-Based Semantic Search and Browsing. *In Proceedings of the First European Semantic Web Symposium (ESWS 2004)*, Heraklion, Greece, 2004.

[HSWW03] Hollink L., Schreiber G., Wielemaker J. et Wielinga B. : Semantic Annotation of Image Collections. *In Workshop on Knowledge Markup and Semantic Annotation, KCAP'03*, Sanibel Island, Florida, USA, 2003.

[Hun03] Hunter J. : Enhancing the Semantic Interoperability of Multimedia through a Core Ontology. *IEEE Transactions on Circuits and Systems for Video Technology, Special*

Issue on Conceptual and Dynamical Aspects of Multimedia Content Description, 13(1), 2003.

[IBL05] ISAAC A., BACHIMONT B. et LAUBLET P. : Indexation de documents audiovisuels : ontologies, patrons de conception et d'utilisation. *In 16ièmes Journées Francophones d'Ingénierie des Connaissances (IC'2005)*, Nice, France, 2005.

[ICG+04] ISAAC A., COUROUNET P., GENEST D., MALAISÉ V., NANARD J. et NANARD M. : Un système d'annotation multiforme et communautaire de documents audiovisuels : Opales. *In Journée sur les Modèles Documentaires de l'Audiovisuel, Semaine du document numérique (SDN 2004)*, La Rochelle, France, June, 22 2004. `http://archivesic.ccsd.cnrs.fr/sic_00001268.html`.

[Isa01] ISAAC A. : Vers la mise en œuvre informatique d'une méthode de conception d'ontologies. Mémoire de D.E.A., Laboratoire LaLICC, Université Paris IV Sorbonne, 2001.

[IT04] ISAAC A. et TRONCY R. : Designing an Audio-Visual Description Core Ontology. *In Workshop on Core Ontologies in Ontology Engineering, 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004)*, Whittlebury Hall, Northamptonshire, UK, 2004. CEUR online Proceedings, `http://ceur-ws.org/Vol-118/`.

[IT05a] ISAAC A. et TRONCY R. : Ontologies et description du contenu de documents AV : une expérimentation dans le domaine médical. *In Atelier Connaissance et Document Temporel, 16ièmes Journées Francophones d'Ingénierie des Connaissances (IC'2005)*, Nice, France, 2005.

[IT05b] ISAAC A. et TRONCY R. : Using Several Ontologies for Describing AV Documents : A Case Study in the Medical Domain. *In Workshop on Multimedia and the Semantic Web, Second European Semantic Web Conference*, Heraklion, Crete, 2005.

[ITM03] ISAAC A., TRONCY R. et MALAISÉ V. : Using XSLT for Interoperability : DOE and The Travelling Domain Experiment. *In 2nd International Workshop on Evaluation of Ontology-based Tools (EON2003) at ISWC2003*, Sanibel Island, Florida, 2003. `http://CEUR-WS.org/Vol-87/`.

[JBV98] JONES D.M., BENCH-CAPON T.J.M. et VISSER P.R.S. : Methodologies for Ontology Development. *In Proc. IT&KNOWS Conference, XV IFIP World Computer Congress*, Budapest, 1998.

[Jou93] JOUIS C. : *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes, Réalisation d'un prototype : le système* SEEK. Thèse de doctorat, EHESS, Paris, 1993.

[KAB+00] KASSEL G., ABEL M.-H., BARRY C., BOULITREAU P., IRASTORZA C. et PERPETTE S. : Construction et exploitation d'une ontologie pour la gestion des connaissances d'une équipe de recherche. *In Actes de la conférence IC'2000 : Journées francophones d'Ingénierie des Connaissances*, Toulouse, 2000.

[Kas99] KASSEL G. : PHYSICIAN is a role played by an object, whereas SIGN is a role played by a concept. *In IJCAI-99 workshop on Ontologies and Problem-Solving Methods*, Stockholm, Sweden, 1999.

[Kas02] KASSEL G. : OntoSpec : une méthode de spécification semi-informelle d'ontologies. *In Actes de la conférence IC'2002 : journées francophones d'Ingénierie des Connaissances*, Rouen, 2002.

[Kay97] Kayser D. : *La représentation des connaissances.* Hermès, Paris, 1997.

[KC95] Kheirbeck A. et Chiaramella Y. : Integrating Hypermedia and Information Retrieval with Conceptual Graphs. *In Proceedings Hypertext-Information Retrieval-Multimedia Conference (HIM'95)*, Konstanz, Germany, 1995.

[KFNM04] Knublauch H., Fergerson R. W., Noy N. F. et Musen M. A. : The Protégé OWL Plugin : An Open Development Environment for Semantic Web Applications. *In Third International Semantic Web Conference (ISWC 2004)*, Hiroshima, Japan, 2004.

[KHGP04] Kalyanpur A., Hashmi N., Golbeck J. et Parsia B. : Lifecycle of a Casual Web Ontology Development Process. *In Proceedings of the WWW2004 Workshop on Application Design, Development and Implementation Issues in the Semantic Web*, New York City, USA, 2004.

[KHPG02] Kalyanpur A., Hendler J., Parsia B. et Golbeck J. : SMORE - Semantic Markup, Ontology, and RDF Editor. Rapport technique, Mindswap, 2002. `http://www.mindswap.org/papers/`.

[Kno04] Consortium KnowledgeWeb : State of the Art on Ontology Alignment. Deliverable 2.2.3, FP6-507482, 2004.

[KP99] Kassel G. et Perpette S. : Co-operative ontology construction needs to carefully articulate terms, notions and objects. *In Ontological Engineering on the Global Information Infrastructure*, Dagstuhl Castle, Germany, 1999.

[Kri82] Kripke S. : *La logique des noms propres.* Seuil, Paris, 1982.

[KV03] Kotis K. et Vouros G. A. : Human Centered Ontology Management in HCONE. *In IJCAI-03 Workshop on Ontologies and Distributed Systems*, Acapulco, Mexico, 2003.

[LA83] Loustalet C. et Association MEDIADOC-Sciences : Décrire l'audiovisuel : manuel méthodologique pour l'analyse de contenu des documents audiovisuels à caractère documentaire. Rapport technique, CNDP, 1983.

[Lak87] Lakoff G. : *Women, Fire and Dangerous Things.* University of Chicago Press, 1987.

[Lal98] Lalmas M. : Logical Models in Information Retrieval : Introduction and Overview. *Information Processing and Management*, 1(34), 1998.

[Le 00] Le Priol F. : *Extraction et capitalisation automatiques de connaissances à partir de documents textuels. SEEK-JAVA : identification et interprétation de relations entre concepts.* Thèse de doctorat, Université Paris IV Sorbonne, 2000.

[Le 03] Le Roux E. : *Extraction d'information dans des textes libres guidée par une ontologie.* Thèse de doctorat, Université Paris X - Nanterre, 2003.

[Len95] Lenat D. B. : Cyc : A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11), 1995.

[Les02] Lespinasse K. : *Acquisition sémantique en langue générale : l'heure de vérité de la paradocumentation textuelle pour l'indexation des documents audiovisuels sur la politique.* Thèse de doctorat, Université Paris III, 2002.

[LH04] Little S. et Hunter J. : Rules-By-Example – a Novel Approach to Semantic Indexing and Querying of Images. *In Proceedings of the Third International Semantic Web Conference, ISWC2004*, Hiroshima, Japan, 2004.

[LHB00] Lespinasse K., Habert B. et Bachimont B. : Le péritexte, un sésame pour les données audiovisuelles ? *In 5èmes Journes Internationales d'Analyse Statistique des Données Textuelles*, Lausanne, Suisse, 2000.

[LN03] Liebig T. et Noppens O. : OntoTrack : Fast Browsing and Easy Editing of Large Ontologies. *In 2nd International Workshop on Evaluation of Ontology-based Tools (EON2003) at ISWC2003*, Sanibel Island, Florida, 2003.

[LWVJ04] Landais R., Wolf C., Vinet L. et Jolion J.-M. : Utilisation de connaissances a priori pour le paramétrage d'un algorithme de détection de textes dans les documents audiovisuels. Appliation à un corpus de journaux télévisés. *In 14ème Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle (RFIA 2004)*, Toulouse, 2004.

[MAH03] Meland P. H., Austvik J. et Heggland J. : Using Ontologies and Semantic Networks with Temporal Media. *In Semantic Web Workshop at the 26th Annual International ACM SIGIR Conference*, Toronto, Canada, 2003.

[Mal05] Malaisé V. : *Méthodologie linguistique et terminologique pour l'aide à la construction d'ontologies différentielles à partir de corpus textuels.* Thèse de doctorat, Université Paris VII, 2005.

[Map95] Maple A. : Faceted Access : A Review of the Literature. Music Library Association Annual Meeting, 1995.

[MBC95] Mechkour M., Berrut C. et Chiaramella Y. : Using Conceptual Graph Framework for Image Retrieval. *In International Conference on Multi-Media Modeling (MMM'95)*, Singapore, 1995.

[MBG+02] Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A. et Schneider L. : WonderWeb Deliverable D17. The WonderWeb Library of Foundational Ontologies and the dolce Ontology. Rapport technique, ISTC-CNR, 2002.

[MCMO03] Martinet J., Chiaramella Y., Mulhem P. et Ounis I. : Photograph indexing and retrieval using star-graphs. *In Third International Workshop on Content-Based Multimedia Indexing (CBMI'03)*, Rennes, France, 2003.

[Men03] Menzel C. : Reference Ontologies – Application Ontologies : Either/Or or Both/And ? *In Workshop on Reference Ontologies vs. Applications Ontologies (26th German Conference on Artificial Intelligence KI 2003)*, Hamburg, Germany, 2003.

[MGLB01] Montes-y-Gomez M., Gelbukh A., Lopez-Lopez A. et Baeza-Yates R. : Flexible Comparison of Conceptual Graphs. *In 12th International Conference on Database and Expert Systems Applications (DEXA 2001)*, Munich, Germany, 2001.

[MGOS04] Mika P., Gangemi A., Oberle D. et Sabou M. : Foundations for Service Ontologies : Aligning OWL-S to DOLCE. *In Proceedings of the Thirteenth International World Wide Web Conference (WWW2004)*, New York, 2004. ACM Press.

[Min81] Minsky M. : A Framework for Representing Knowledge. *In* Haugeland J., éditeur : *Mind Design.* MIT Press, 1981.

[Mou99] Moulis A.-M. : L'indexation de la collection « Cinéastes de notre temps ». *In L'indexation à l'ère d'Internet. Colloque d'ISKO-France*, Lyon, France, 1999.

[mpeg-701] mpeg-7 : Information Technology – Multimedia Content Interface. Rapport technique 15938, ISO/IEC, 2001.

[MSD00] Motta E., Shum S. B. et Domingue J. : Ontology-driven document enrichment : principles, tools and applications. *International Journal of Human Computer Studies*, 52(6), 2000.

[MSSS00] A. Maedche, H.-P. Schnurr, S. Staab et R. Studer : Representation Language-Neutral Modeling of Ontologies. *In Proceedings of the German Workshop "Modellierung"*, Koblenz, Germany, 2000.

[MZB04] Malaisé V., Zweigenbaum P. et Bachimont B. : Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologies. *In 11e édition de la Conférence sur le Traitement Automatique du Langage Naturel (TALN 2004)*, Fès, Maroc, 2004.

[NB96] Nie J.-Y. et Brisebois M. : An Inferential Approach to Information Retrieval and its Implementation using a Manual Thesaurus. *Artificial Intelligence Review*, 10(5-6), 1996.

[New82] Newell A. : The knowledge level. *Artificial Intelligence*, 18(1), 1982.

[NFM00] Noy N.F., Fergerson R. W. et Musen M. A. : The knowledge model of Protege-2000 : Combining interoperability and flexibility. *In 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, Juan-les-Pins, France, 2000.

[NM01] Noy N.F. et McGuinnes, D.L. : Ontology Development 101 : A Guide to Creating Your First Ontology. Rapport technique SMI-2001-0880, SMI, 2001.

[NM03] Noy N. F. et Musen M. A. : The PROMPT suite : interactive tools for ontology merging and mapping. *International Journal of Human-Computer-Studies*, 59(6), 2003.

[NNK03] Nanard M., Nanard J. et King P. : IUHM, a Hypermedia-based Model for Integrating Open Services, Data and Metadata. *In International ACM Conference on Hypertext (HT'2003)*, Nottingham, UK, 2003.

[OGGM02] Oltramari A., Gangemi A., Guarino N. et Masolo C. : Restructuring WordNet's Top-Level : The OntoClean approach. *In LREC02 Workshop on Ontologies and Lexical Knowledge Bases (OntoLex 2002)*, Las Palmas, Canary Islands, Spain, 2002.

[OP98] Ounis I. et Pasca M. : Modeling, Indexing and Retrieving Images using Conceptual Graphs. *In 9th International Conference on Database and Expert Systems Applications (DEXA'98)*, 1998.

[Oun98] Ounis I. : *Un modèle d'indexation relationnel pour les graphes conceptuels fondé sur une interprétation logique.* Thèse de doctorat, Université Joseph Fourier, 1998.

[OWL04] OWL : Web Ontology Language Reference. W3C Recommendation, 2004. `http://www.w3.org/TR/owl-ref/`.

[Pa03] Pool M. et alii. : Evaluating Expert-Authored Rules for Military Reasoning. *In Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP'03)*, Sanibel Island, Florida, 2003.

[PC05] Poli J.-P. et Carrive J. : Proposition d'une architecture pour un système de structuration automatique de flux audiovisuels. *In Actes de la Conférence sur la COmpression et la REprésentation des Signaux Audiovisuels 2005 (CORESA'05)*, Rennes, 2005.

[Péd03] R T. Pédauque : Le document : forme, signe et relation, les re-formulations du numérique. Document de Travail, Réseau Thématique Pluridisciplinaire 33, Département STIC, CNRS, 2003. `http://archivesic.ccsd.cnrs.fr/sic_00000511.html`.

[Pic96] Pichon J. : Le traitement documentaire des programmes de radio et de télévision. Rapport technique, INA, 1996.

[Pus01] Pustejowsky J. : Type Construction and the Logic of Concepts. *In* Bouillon P. et Busa F., éditeurs : *The Syntax of Word Meaning*. Cambridge University Press, 2001.

[QKH03] Quan D., Karger D.R. et Huynh D. F. : RDF Authoring Environments for End Users. *In International Workshop on Semantic Web Foundations and Application Technologies (SWFAT-2003)*, Nara, Japan, 2003.

[Qui68] Quillian M. R. : Semantic Memory. *In Semantic Information Processing*. MIT Press, 1968.

[Ran00] Ranwez S. : *Composition Automatique de Documents Hypermédia Adaptatifs à partir d'Ontologies et de Requêtes Intentionnelles de l'Utilisateur.* Thèse de doctorat, Université Montpellier II, 2000.

[RBF+02] Rousset M.-C., Bidault A., Froidevaux C., Gagliardi H., Goasdoué F., Reynaud C. et Safar B. : Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL. *Revue Information – Interaction – Intelligence*, 2(1), 2002.

[RBv+00] Rutledge L., Bailey B., van Ossenbruggen J., Hardman L. et Geurts J. : Generating Presentation Constraints from Rhetorical Structure. *In Proceedings of the 11th ACM conference on Hypertext and Hypermedia*, San Antonio, Texas, 2000.

[RCA94] Rastier F., Cavazza M. et Abeillé A. : *Sémantique pour l'Analyse*. Masson, Paris, 1994.

[RCP02] Roussey C., Calabretto S. et Pinon J.-M. : Le thésaurus sémantique : contribution à l'ingénierie des connaissances documentaires. *In Actes de la conférence IC'2002 : journées francophones d'Ingénierie des Connaissances*, Rouen, 2002.

[RDF04a] RDF : Ressource Description Framework Primer. W3C Recommendation, 2004. http://www.w3.org/TR/rdf-primer/.

[RDF04b] RDFS : RDF Vocabulary Description Language : RDF Schema. W3C Recommendation, 2004. http://www.w3.org/TR/rdf-schema/.

[RDH+04] Rector A., Drumond N., Horridge J., Rogers J., Knublauch H., Stevens R., Wang H. et Wroe C. : owl Pizzas : Practical Experience of Teaching owl-dl : Common Errors & Common Patterns. *In 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2004)*, Northamptonshire, UK, 2004.

[Rec03] Rector A. : Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including owl. *In Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP'03)*, Sanibel Island, Florida, 2003.

[Rei99] Reich J. R. : Ontological Design Patterns for the Integration of Molecular Biological Information. *In German Conference on Bioinformatics (GCB 99)*, Hannover, Germany, 1999.

[RGSF99] Rieu D., Giraudin J.-P., Saint-Marcel C. et Front-Conte A. : Des opérations et des relations pour les patrons de conception. *In Congrès Inforsid'99*, Toulon, 1999.

[RZS+99] Rector A., Zanstra P.E., Solomon W. D., Rogers J. E., Baud R., Ceusters W., Claassen W., Kirby J., Rodrigues J.-M., Rossi Mori A., van der Haring E. J. et Wagner J. : Reconciling User's Needs and Formal Requirements : Issues in developping a Re-Usable Ontology for Medecine. *IEEE Transactions on Information Technology in BioMedecine*, 4(2), 1999.

[SAA+99] Schreiber G., Akkermans H., Anjewierden A., de Hoog R., Shadbolt N., van de Velde W. et Wielinga B. : *Knowledge Engineering and Management. The CommonKADS Methodology.* MIT Press, Cambridge, Massachusetts, 1999.

[Sal97] Salvat E. : *Raisonner avec des opérations de graphes : Graphes conceptuels et règles d'inférences.* Thèse de doctorat, Université Montpellier II, 1997.

[SBC+02] Schreiber A. T., Blok I., Carlier D., van Gent W., Hokstam J. et Roos U. : A Mini-Experiment in Semantic Annotation. *In International Semantic Web Conference (ISWC 2002)*, Sardinia, Italia, 2002.

[SCB+04] Saggion H., Cunningham H., Bontcheva K., Maynard D., Hamza O. et Wilks Y. : Multimedia indexing through multi-source and multi-language information extraction : the MUMIS project. *Data and Knowledge Engineering*, 48(2), 2004.

[SCC03] Stockinger P., Challulau H. et Courounet P. : Evaluation de l'environnement Opales. Rapport technique Rapport WP5 du projet Opales, Maison des Sciences de l'Homme, 2003.

[SDWW01] Schreiber A. Th., Dubbeldam B., Wielemaker J. et Wielinga B. : Ontology-Based Photo Annotation. *IEEE Intelligent Systems*, 16(3), 2001.

[SEA+02] Sure Y., Erdmann M., Angele J., Staab S., Studer R. et Wenke D. : OntoEdit : Collaborative Ontology Development for the Semantic Web. *In International Semantic Web Conference (ISWC 2002)*, Sardinia, Italia, 2002.

[SEM01] Staab S., Erdmann M. et Maedche A. : Engineering Ontologies Using Semantic Patterns. *In IJCAI-2001 Workshop on E-Business and Intelligent Web*, Seattle, USA, 2001.

[SJ02] Simov K. et Jordanov S. : BOR : a Pragmatic DAML+OIL Reasoner. Rapport technique 40, On-To-Knowledge Project, 2002.

[SLL+04] Soergel D., Lauser B., Liang A., Fisseha F., Keizer J. et Katz S. : Reengineering Thesauri for New Applications : the Agrovoc Example. *Journal of Digital Information*, 4(4), 2004.

[SM00] Staab S. et Maedche A. : Ontology Engineering beyond the modeling of concepts and relations. *In Proceedings of the ECAI-2000 Workshop on Ontologies and Problem-Solving Methods*, Berlin, 2000.

[Sow84] Sowa J. F. : *Conceptual structures : information processing in mind and machine.* Addisson-Wesley, Reading (MA US), 1984.

[Sow00] Sowa J. F. : *Knowledge Representation : Logical, Philosophical, and Computational Foundations.* Brooks Cole, Pacific Grove, California, 2000.

[Sto03a] Stockinger P. : *Le document audiovisuel – Procédures de description et exploitation.* Hermès Science, Paris, France, 2003.

[Sto03b] Stockinger P. : Opales : Digital libraries for humanities. *In Workshop on Multimedia in Digital Libraries*, 2003.

[Svd+04] Stuckenschmidt H., van Harmelen F., de Waard A., Scerri T., Bhogal R., van Buel J., Crowlesmith I., Fluit C., Kampman A., Broekstra J. et van Mulligen E. : Exploring Large Document repositories with RDF Technology : the DOPE Project. *IEEE Intelligent Systems*, 19(3), 2004.

[TAJ01] Tudhope D., Alani H. et Jones C. : Augmenting Thesaurus Relationships : Possibilities for Retrieval. *Journal of Digital Information*, 1(8), 2001.

[Tam02] Tamma V.A.M. : *An Ontology Model supporting Multiple Ontologies for Knowledge Sharing*. Thèse de doctorat, University of Liverpool, 2002.

[TBC+02] Thomere J., Barker K., Chaudhri V., Clark P., Eriksen M., Mishra S., Porter B. et Rodriguez A. : Web-based Ontology Browsing and Editing System. *In Proceedings of the 14th Innovative Applications of AI Conference (IAAI 02)*, Edmonton, Alberta, Canada, 2002.

[TBH03] Thomopoulos R., Buche P. et Haemmerlé O. : Du flou dans les graphes conceptuels. *In 3èmes Journées Nationales sur les Modèles de Raisonnement (JNMR'03)*, Paris, 2003.

[TI02] Raphaël Troncy et Antoine Isaac : DOE : une mise en oeuvre d'une méthode de structuration différentielle pour les ontologies. *In 13th Journées Francophones d'Ingénierie des Connaissances (IC'02)*, pages 63–74, Rouen, France, May, 28-30 2002.

[TNL+05] Tolksdorf R., Nixon L. J. B., Liebsch F., Minh Nguyen D., Paslaru Bontas E. et Nixon L. J. B. : Enabling real world Semantic Web applications through a coordination middleware. *In Proceedings of the 2nd European Semantic Web Conference (ESWC 2005)*, Heraklion, Greece, 2005.

[TPC04] Tsinaraki C., Polydoros P. et Christodoulakis S. : Integration of OWL ontologies in MPEG-7 and TV-Anytime compliant Semantic Indexing. *In 16th International Conference on Advanced Information Systems Engineering (CAiSE 2004)*, Riga, Latvia, 2004.

[Tro04] Troncy R. : *Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies. Application à la description de documents audiovisuels traitant du cyclisme*. Thèse de doctorat, Université Joseph Fourier - Grenoble I, 2004.

[Tur98] Turner J. : *Images en mouvements. Stockage – Repérage – Indexation*. Presses de l'Université du Québec, Québec, Canada, 1998.

[TV03] Tempich C. et Volz R. : Towards a benchmark for Semantic Web reasoners – An analysis of the DAML ontology library. *In 2nd International Workshop on Evaluation of Ontology-based Tools (EON2003) at ISWC2003*, Sanibel Island, Florida, 2003.

[TVV04] Thièvre J., Viaud M.-L. et Verroust-Blondet A. : Using Euler Diagrams in Traditional Library Environments. *In First International Workshop on Euler Diagrams (Euler 2004)*, Brighton, United Kingdom, 2004.

[UG96] Uschold M. et Gruninger M. : Ontologies : Principles, Methods and Applications. *Knowledge Engineering Review*, 11(2), 1996.

[UK95] Uschold M. et King M. : Towards a Methodology for Building Ontologies. *In IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Montréal, Canada, 1995.

[van86] van Rijsbergen C. J. : A new theoretical framework for information retrieval. *In ACM Conference on Research and development in Information Retrieval*, Pisa, Italy, 1986.

[VB96] Valente A. et Breuker J. : Towards Principled Core Ontologies. *In Proceedings of the Knowledge Acquisition Workshop (KAW'96)*, Banff, Canada, 1996.

[Ven02] Veneau E. : *Macro-segmentation multi-critère et classification des séquences par le contenu dynamique pour l'indexation vidéo.* Thèse de doctorat, Université de Rennes I, 2002.

[vMS+04] van Assem M., Menken M. R., Schreiber G., Wielemaker J. et Wielinga B. : A method for converting thesauri to RDF/OWL. *In 3rd International Semantic Web Conference (ISWC 2004)*, 2004.

[VPS03] Vakkari P., Pennamen M. et Serola S. : Changes of search terms and tactics while writing a research proposal : A longitudinal cas study. *Information Processing and Management*, 39(3), 2003.

[Wag03] Wagner G. : Seven Golden Rules for a Web Rule Language. *IEEE Intelligent Systems*, 5(18), 2003.

[Wal99] Waller S. : *L'analyse documentaire. Une approche méthodologique.* ADBS, Paris, 1999.

[WG01] Welty C. et Guarino N. : Supporting Ontological Analysis of Taxonomic Relationships. *Data and Knowledge Engineering*, 1(39), 2001.

[WSWS01] Wielinga B., Schreiber A. Th., Wielemaker J. et Sandberg J. A. C. : From Thesaurus to Ontology. *In Proceedings of the 1st International Conference on Knowledge Capture (K-Cap'01)*, Victoria, British Columbia, Canada, 2001.

[XML04] XML : Extensible Markup Language Version 1.0, Third Edition. W3C Recommendation, 2004. `http://www.w3.org/TR/REC-xml/`.

[XSL99] XSLT : XSL Transformations. W3C Recommendation, 1999. `http://www.w3.org/TR/xslt`.

[Zar00] Zarri G. P. : A Conceptual Model for Implementing Metadata on the Web. *In Fifth conference on Current Research Information Systems, CRIS 2000*, Helsinki, Finland, 2000.

[ZBB+99] Zarri G. P., Bertino E., Black B., Brasher A., Catania B., Deavin D., Di Pace L., Esposito F., Leo P., McNaught J., Persidis A., Rinaldi F. et Semeraro G. : CONCERTO, An Environment for the "Intelligent" Indexing, Querying and Retrieval of Digital Documents. *In LNAI 1609 : Foundations of Intelligent Systems, Proc. of the 11th Int. Symp. ISMIS'99*, Warsaw, Poland, 1999.

[ZC94] Zweigenbaum P. et Consortium Menelas : Menelas : an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, 45(1-2), 1994.

[Zet02] Zettl H. : Television Aesthetics. *In* Adler R. P., éditeur : *Understanding Television.* Praeger, New York, 2002.

Résumé

L'indexation de documents audiovisuels est une tâche difficile mais nécessaire si l'on veut rechercher correctement des contenus dans des bases volumineuses. Pour atteindre ses objectifs, un système d'indexation doit imposer un certain degré de contrôle, et assister ses utilisateurs tant pour la création que pour la recherche des index. Dans ce manuscrit, nous discutons de la façon dont on peut concevoir des systèmes d'indexation et de recherche à base de connaissances (SBC). En les adaptant aux pratiques de l'INA et aux besoins de projets expérimentaux comme OPALES, nous reprenons les concepts et techniques de la représentation des connaissances et du web sémantique. Nous nous concentrons principalement sur les ontologies, modèles conceptuels formalisés sur lesquels s'appuient les SBC pour le contrôle des descriptions et la conduite de raisonnements. Nous mettons en particulier en œuvre une méthode d'assistance à la conception des ontologies, et insistons sur l'intérêt d'utiliser des patrons de conception ontologiques. Dans nos réflexions figure aussi la notion de patrons d'indexation, structures relationnelles reflétant les index typiques d'une application donnée. Les patrons d'indexation facilitent la création des index, tout comme ils permettent de rationaliser la conception des ontologies, notamment celle des connaissances de raisonnement qui spécifient les inférences réalisées par le SBC. Nos propositions méthodologiques ont été appliquées dans un certain nombre de réalisations informatiques et d'expérimentations pour des domaines réels, que nous évoquons et discutons tout au long de ce manuscrit.
Discipline: Informatique
Mots-clés: Ontologies, Indexation sémantique, Indexation de documents audiovisuels, Construction d'ontologies, Patrons de conception, Patrons d'indexation, Raisonnement

Abstract

Indexing audiovisual documents is a tedious work. However, it is needed if one wants to search efficiently for specific contents inside a large document base. To reach its goal, an indexing system has to ensure proper control and assistance with respect to the creation and the search for indices. In this thesis, we discuss the way knowledge-based indexing systems ought to be designed. We turn to the notions and tools introduced in the knowledge representation field and the Semantic Web initiative, and try to adapt them to the practices found in INA, as well as to the special needs of experimental projects such as OPALES. We mainly focus on ontologies, which are formalised conceptual models that can be used to specify the abilities of knowledge-based systems regarding the control of descriptions and the reasoning processes. We especially implement a method assisting ontology creation, and insist on the use of ontological design patterns. Central to our thoughts also lies the notion of indexing patterns. Those relational structures mirror the typical indices that can be found in a given application, and therefore help index creation. Additionally, they might be used to rationalise ontology conception, especially regarding the design of the formal reasoning knowledge which specifies the inferences a knowledge-based system can achieve. Our methodological proposals have been applied in a range of concrete implementations and real-domain experiments we detail and discuss throughout this thesis.
English Title: Building and Using Ontologies for Audiovisual Document Indexing